RESEARCH CENTRE

Sophia Antipolis - Méditerranée

IN PARTNERSHIP WITH:

Université Côte d'Azur

2020 ACTIVITY REPORT

Project-Team MAASAI

Models and Algorithms for Artificial Intelligence

IN COLLABORATION WITH: Laboratoire Jean-Alexandre Dieudonné (JAD), Laboratoire informatique, signaux systèmes de Sophia Antipolis (I3S)

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Contents

Project-Team MAASAI

Creation of the Project-Team: 2020 February 01

Keywords

Computer sciences and digital sciences

A3.1. – Data

A3.1.10. – Heterogeneous data

A3.1.11. – Structured data

A3.4. – Machine learning and statistics

A3.4.1. – Supervised learning

A3.4.2. – Unsupervised learning

A3.4.6. – Neural networks

A3.4.7. – Kernel methods

A3.4.8. – Deep learning

A9. – Artificial intelligence

A9.2. – Machine learning

Other research topics and application domains

B6.3.4. – Social Networks

B7.2.1. – Smart vehicles

B8.2. – Connected city

B9.6. – Humanities

1 Team members, visitors, external collaborators

Research Scientist

- Pierre Alexandre Mattei [Inria, Researcher, from Feb 2020]

Faculty Members

- Charles Bouveyron [Team leader, Univ Côte d'Azur, Professor, from Feb 2020, HDR]
- Marco Corneli [Univ Côte d'Azur, from Feb 2020]
- Damien Garreau [Univ Côte d'Azur, Associate Professor, from Feb 2020]
- Frederic Precioso [Univ Côte d'Azur, Professor, from Feb 2020]
- Michel Riveill [Univ Côte d'Azur, Professor, from Feb 2020]

Post-Doctoral Fellows

- Juliette Chevallier [Univ Côte d'Azur, from Feb 2020]
- Gabriel Wallin [Univ Côte d'Azur, from Feb 2020]

PhD Students

- Yassine El Amraoui [Ezako, CIFRE, From September 2020]
- Edson Florez [R & France Lab, From February 2017]
- John Anderson Garcia Henao [Unive Côte d'Azur, From September 2017]
- Laurent Garcia Henao [Univ Côte d'Azur, from Oct 2018]
- Dingge Liang [Univ Côte d'Azur]
- Giulia Marchello [Univ Côte d'Azur, from Sep 2020]
- Taki Eddine Mekhalfa [Univ Côte d'Azur, from Oct 2020]
- Hugo Miralles [Orange, CIFRE, from Dec 2020]
- Baptiste Pouthier [NXP, CIFRE, from Nov 2020]
- Miguel Romero Rondon [Univ Côte d'Azur, ATER, from Oct 2017]
- Laura Sanabria Rosas [Universite Cote d'Azur, ATER, from Oct 2017]
- Hugo Schmutz [Univ Côte d'Azur, from Oct 2020]
- Cedric Vincent-Cuaz [Univ Côte d'Azur, From Nov 2020]
- Tianshu YANG [Amadeus, CIFRE, from October 2017]
- Xuchun Zhang [Univ Côte d'Azur, from Sep 2020]
- Mansour Zoubeirou A Mayaki [Pro BTP, CIFRE, from Dec 2020]

Interns and Apprentices

- Gustavo Gavanzo Chaves [Univ Côte d'Azur, from Apr 2020 until Aug 2020]
- Onkar Jadhav [Univ Côte d'Azur, from Apr 2020 until Aug 2020]
- Gayathri Kotte [Univ Côte d'Azur, from Apr 2020 until Aug 2020]
- Giulia Marchello [Univ Côte d'Azur, until May 2020]
- Lola Morin [Ecole Supérieure des Technologies Industrielles Avancées, from Feb 2020 until Jun 2020]
- Louis Ohl [INSA Lyon, from Sep 2020]

Administrative Assistant

- Nathalie Brillouet [Inria, from Feb 2020]

External Collaborators

- Elena Erosheva [University of Washington & Univ Côte d'Azur, from Mar 2020, HDR]
- Marco Gori [Universita di Sienna & Institut 3IA Côte d'Azur, HDR]

2 Overall objectives

Artificial intelligence has become a key element in most scientific fields and is now part of everyone life thanks to the digital revolution. Statistical, machine and deep learning methods are involved in most scientific applications where a decision has to be made, such as medical diagnosis, autonomous vehicles or text analysis. The recent and highly publicized results of artificial intelligence should not hide the remaining and new problems posed by modern data. Indeed, despite the recent improvements due to deep learning, the nature of modern data has brought new specific issues. For instance, learning with high-dimensional, atypical (networks, functions, ...), dynamic, or heterogeneous data remains difficult for theoretical and algorithmic reasons. The recent establishment of deep learning has also opened new questions such as: How to learn in an unsupervised or weakly-supervised context with deep architectures? How to design a deep architecture for a given situation? How to learn with evolving and corrupted data?

To address these questions, the Maasai team focuses on topics such as unsupervised learning, theory of deep learning, adaptive and robust learning, and learning with high-dimensional or heterogeneous data. The Maasai team conducts a research that links practical problems, that may come from industry or other scientific fields, with the theoretical aspects of Mathematics and Computer Science. In this spirit, the Maasai project-team is totally aligned with the "Core elements of AI" axis of the Institut 3IA Côte d'Azur. It is worth noticing that the team hosts two 3IA chairs of the Institut 3IA Côte d'Azur, as well as several PhD students funded by the Institut.

3 Research program

Within the research strategy explained above, the Maasai project-team aims at developing statistical, machine and deep learning methodologies and algorithms to address the following four axes.

Unsupervised learning The first research axis is about the development of models and algorithms designed for unsupervised learning with modern data. Let us recall that unsupervised learning — the task of learning without annotations — is one of the most challenging learning challenges. Indeed, if supervised learning has seen emerging powerful methods in the last decade, their requirement for huge annotated data sets remains an obstacle for their extension to new domains. In addition, the nature of modern data significantly differs from usual quantitative or categorical data. We ambition in this

axis to propose models and methods explicitly designed for unsupervised learning on data such as high-dimensional, functional, dynamic or network data. All these types of data are massively available nowadays in everyday life (omics data, smart cities, ...) and they remain unfortunately difficult to handle efficiently for theoretical and algorithmic reasons. The dynamic nature of the studied phenomena is also a key point in the design of reliable algorithms.

On the one hand, we direct our efforts towards the development of unsupervised learning methods (clustering, dimension reduction) designed for specific data types: high-dimensional, functional, dynamic, text or network data. Indeed, even though those kinds of data are more and more present in every scientific and industrial domains, there is a lack of sound models and algorithms to learn in an unsupervised context from such data. To this end, we have to face problems that are specific to each data type: How to overcome the curse of dimensionality for high-dimensional data? How to handle multivariate functional data / time series? How to handle the activity length of dynamic networks? On the basis of our recent results, we ambition to develop generative models for such situations, allowing the modeling and the unsupervised learning from such modern data.

On the other hand, we focus on deep generative models (statistical models based on neural networks) for clustering and semi-supervised classification. Neural network approaches have demonstrated their efficiency in many supervised learning situations and it is of great interest to be able to use them in unsupervised situations. Unfortunately, the transfer of neural network approaches to the unsupervised context is made difficult by the huge amount of model parameters to fit and the absence of objective quantity to optimize in this case. We therefore study and design model-based deep learning methods that can hande unsupervised or semi-supervised problems in a statistically grounded way.

Finally, we also aim at developing explainable unsupervised models that can ease the interaction with the practitioners and their understanding of the results. There is an important need for such models, in particular when working with high-dimensional or text data. Indeed, unsupervised methods, such as clustering or dimension reduction, are widely used in application fields such as medicine, biology or digital humanities. In all these contexts, practitioners are in demand of efficient learning methods which can help them to make good decisions while understanding the studied phenomenon. To this end, we aim at proposing generative and deep models that encode parsimonious priors, allowing in turn an improved understanding of the results.

Understanding (deep) learning models The second research axis is more theoretical, and aims at improving our understanding of the behaviour of modern machine learning models (including, but not limited to, deep neural networks). Although deep learning methods and other complex machine learning models are obviously at the heart of artificial intelligence, they clearly suffer from an overall weak knowledge of their behaviour, leading to a general lack of understanding of their properties. These issues are barriers to the wide acceptance of the use of AI in sensitive applications, such as medicine, transportation, or defense. We aim at combining statistical (generative) models with deep learning algorithms to justify existing results, and allow a better understanding of their performances and their limitations.

We particularly focus on researching ways to understand, interpret, and possibly explain the predictions of modern, complex machine learning models. We both aim at studying the empirical and theoretical properties of existing techniques (like the popular LIME), and at developing new frameworks for interpretable machine learning (for example based on deconvolutions or generative models). Among the relevant application domains in this context, we focus notably on text and biological data.

Another question of interest is: what are the statistical properties of deep learning models and algorithms? Our goal is to provide a statistical perspective on the architectures, algorithms, loss functions and heuristics used in deep learning. Such a perspective can reveal potential issues in exisiting deep learning techniques, such as biases or miscalibration. Consequently, we are also interested in developing statistically principled deep learning architectures and algorithms, which can be particularly useful in situations where limited supervision is available, and when accurate modelling of uncertainties is desirable.

Adaptive and Robust Learning The third research axis aims at designing new learning algorithms which can learn incrementally, adapt to new data and/or new context, while providing predictions robust to

biases even if the training set is small.

For instance, we have designed an innovative method of so-called cumulative learning, which allows to learn a convolutional representation of data when the learning set is (very) small. The principle is to extend the principle of Transfer Learning, by not only training a model on one domain to transfer it once to another domain (possibly with a fine-tuning phase), but to repeat this process for as many domains as available. We have evaluated our method on mass spectrometry data for cancer detection. The difficulty of acquiring spectra does not allow to produce sufficient volumes of data to benefit from the power of deep learning. Thanks to cumulative learning, small numbers of spectra acquired for different types of cancer, on different organs of different species, all together contribute to the learning of a deep representation that allows to obtain unequalled results from the available data on the detection of the targeted cancers. This extension of the well-known Transfer Learning technique can be applied to any kind of data.

We also investigate active learning techniques. We have for example proposed an active learning method for deep networks based on adversarial attacks. An unlabelled sample which becomes an adversarial example under the smallest perturbations is selected as a good candidate by our active learning strategy. This does not only allow to train incrementally the network but also makes it robust to the attacks chosen for the active learning process.

Finally, we address the problem of biases for deep networks by combining domain adaptation approaches with Out-Of-Distribution detection techniques.

Learning with heterogeneous and corrupted data The last research axis is devoted to making machine learning models more suitable for real-world, "dirty" data. Real-world data rarely consist in a single kind of Euclidean features, and are genereally heterogeneous. Moreover, it is common to find some form of corruption in rea-world data sets: for example missing values, outliers, label noise, or even adversarial examples.

Heterogeneous and non-Euclidean data are indeed part of the most important and sensitive applications of artificial intelligence. As a concrete example, in medicine, the data recorded on a patient in an hospital range from images to functional data and networks. It is obviously of great interest to be able to account for all data available on the patients to propose a diagnostic and an appropriate treatment. Notice that this also applies to autonomous cars, digital humanities and biology. Proposing unified models for heterogeneous data is an ambitious task, but first attempts (e.g. the Linkage[1] project) on combination of two data types have shown that more general models are feasible and significantly improve the performances. We also address the problem of conciliating structured and non-structured data, as well as data of different levels (individual and contextual data).

On the basis of our previous works (notably on the modeling of networks and texts), we first intend to continue to propose generative models for (at least two) different types of data. Among the target data types for which we would like to propose generative models, we can cite images and biological data, networks and images, images and texts, and texts and ordinal data. To this end, we explore modelings trough common latent spaces or by hybridizing several generative models within a global framework. We are also interested in including potential corruption processes into these heterogeneous generative models. For example, we are developping new models that can handle missing values, under various sorts of missingness assumptions.

Besides the modelling point of view, we are also interested in making existing algorithms and implementations more fit for "dirty data". We study in particular ways to robustify algorithms, or to improve heuristics that handle missing/corrupted values or non-Euclidean features.

4 Application domains

The Maasai research team has the following major application domains:

Medicine Most of team members apply their research work to Medicine or extract theoretical AI problems from medical situations. In particular, our main applications to Medicine are concerned with

[1] The Linkage project: https://linkage.fr

pharmacovigilance, medical imaging, and omics. It is worth noticing that medical applications cover all research axes of the team due to the high diversity of data types and AI questions. It is therefore a preferential field of application of the models and algorithms developed by the team.

Digital humanities Another important application field for Maasai is the increasingly dynamic one of digital humanities. It is a extremly motivating field due to the very orginal questions that are addressed. Indeed, linguists, sociologists, geographers and historians have questions that are quite different than the usual ones in AI. This allows the team to formalize original AI problems that can be generalized to other fields, allowing to inderectly contribute to the general theory and methodology of AI.

Computer Vision The last main application domain for Maasai is computer vision. With the revolution brought to computer vision field by deep learning techniques, new questions have appeared such as combining subsymbolic and symbolic approaches for complex semantic and perception problems, or as edge AI to embed machine learning approaches for computer vision solutions preserving privacy. This domain brings new AI problems which require to bridge the gap between different views of AI.

Other application domains Other topics of interest of the team include astronomy, bioinformatics, and recommender systems.

5 Highlights of the year

The team has been created in 2020.

5.1 Awards and chairs

The team has two 3IA chairs of the Institut 3IA Côte d'Azur:

- Charles Bouveyron, 3IA chair
- Marco Gori, 3IA international chair

These two chairs were selected by the international jury of the national 3IA program in 2019.

5.2 Conference organisation

Pierre-Alexandre Mattei co-founded and co-organized the first ICML workshop on the Art of learning with missing values (Artemiss). Artemiss was part of the 2020 edition of the International Conference on machine Learning (ICML), the premier machine learning conference. It was the first workshop on missing data included in one of the major machine learning conference in recent years. The other founding co-organisers are Jes Frellsen (Technical University of Denmark), Julie Josse (Inria Sophia-Antipolis), and Gaël Varoquaux (Inria Saclay).
Workshop website: https://artemiss-workshop.github.io/

5.3 Publication of reference books

Chales Bouveyron has published the book "Model-Based Clustering and Classification for Data Science" with Adrian Raftery (University of Washington, USA), Brendan Murphy (University College Dublin, Ireland) and Gilles Celeux (Inria Saclay), in the "Series in Statistical and Probabilistic Mathematics" of Cambridge University Press. Charles Bouveyron ensures the maintenance of the book website, where a free PDF version of the book is available, as well as all the R scripts of the book.
Book website: https://math.unice.fr/~cbouveyr/MBCbook/

6 New software and platforms

For the Maasai research team, the main objective of the software implementations is to experimentally validate the results obtained and ease the transfer of the developed methodologies to industry. Most of the software will be released as R or Python packages that requires only a light maintaining, allowing a relative longevity of the codes. Some platforms are also proposed to ease the use of the developed methodologies by users without a strong background in Machine Learning, such as scientists from other fields.

6.1 Softwares

6.1.1 R packages

- OrdinalLBM: https://cran.r-project.org/web/packages/ordinalLBM/index.html
- FunLBM: https://cran.r-project.org/web/packages/funLBM/index.html
- SpinyReg: https://cran.r-project.org/web/packages/spinyReg/index.html

6.1.2 Python

- HDMI: https://gist.github.com/ahoudard/677943abbfc7308d30fbbe749d353d68
- MIWAE: https://github.com/pamattei/miwae
- not-MIWAE: https://github.com/nbip/notMIWAE

6.1.3 Julia

- PEN: https://github.com/SamuelWiqvist/PENs-and-ABC

6.2 Platforms

6.2.1 Linkage: a statistical AI algorithm to analyze communication networks

Platform address: https://linkage.fr
Participants: Charles Bouveyron, Marco Corneli
Keywords: generative models, clustering of networks, communication networks, publication networks, cyberdefense
Collaborations: Pierre Latouche (Univ. de Paris)

Even though the clustering field has received strong attention, the joint analysis of texts and networks has been very limited while most social networks are based on texts (emails, Facebook, Twitter, ...). The implemented methodology used in the software Linkage tackles this limit and brings a solution so that texts and networks are analyzed simultaneously. Linkage implements a statistical model along with an AI algorithm for the inference which enables the segmentation of the nodes (individuals) of a network with textual edges, while identifying the topics of discussion. The networks can be directed (emails, tweets, ...) or undirected (scientific papers, posts on Facebook, ...). Linkage only requires the data of a set of textual exchanges between people or more generally between entities. For instance, we may consider the exchanges of texts between individuals on a social network, or the exchanges of emails between employees from the same company, or even the co-publications of patents or scientific papers. Linkage is also able to automatically determine the number of groups of individuals and the number of topics used. Thus, data processing is completely automatic. The outputs of the software are the classification of individuals in groups (as well as probabilities) and a description of topics with the most specific words. A demonstration platform is available at the following address: https://linkage.fr. It allows anyone wishing to, to use Linkage with their own data or with public data. Registration is free on the website but the volume of data to be treated by the platform is limited. The best computing libraries and the newest security protocols are used to guarantee an optimal treatment of data. The platform offers different

ways of uploading data. First, it is possible to upload personal data in a simple format (.csv), but also to collect personal email networks via our Gmail App and finally, it is possible to run public queries (on Twitter, PubMed, Arxiv, ...). There are many graphic layouts available on the platform to facilitate the use of the results. To finish, if needed, raw data can be downloaded to run them afterwards in a visualization software such as Gephi for example.

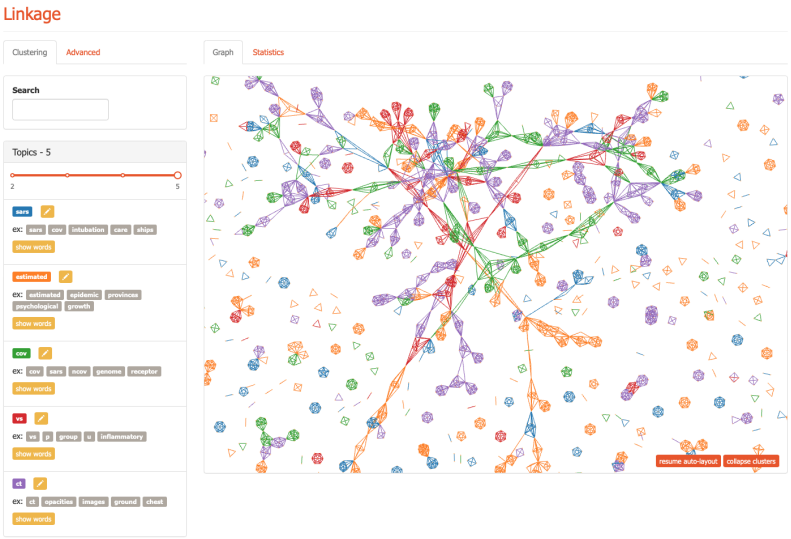


Figure 1: The Linkage.fr platform which implements and allows to interact with the STBM model.

6.2.2 Topix: a AI-based solution allowing to summarize massive and sparse text data

Platform address: https://topix.mi.parisdescartes.fr
Participants: Charles Bouveyron, Marco Corneli
Keywords: generative models, model-based co-clustering, massive and sparse text data
Collaborations: Pierre Latouche (Univ. de Paris), Laurent Bergé (Univ. Luxembourg)

Topix is an innovative AI-based solution allowing to summarize massive and possibly extremely sparse data bases involving text. Topix is a versatile technology that can be applied in a large variety of situations where large matrices of texts / comments / reviews are written by users on products or addressed to other individuals (bi-partite networks). The typical use case consists in a e-commerce company interested in understanding the relationship between its users and the sold products thanks to the analysis of user comments. A simultaneous clustering (co-clustering) of users and products is produced by the Topix software, based on the key emerging topics from the reviews and by the underlying model. The Topix demonstration platform allows to upload the user data on the website, in a totally secured framework, and let the AI-based software analyze them for the user. The platform also proposes some typical use cases to give a better idea of what Topix can do.

6.2.3 DiagnoseNET: Automatic Framework to Scale Neural Networks on Heterogeneous Systems

Platform address: https://diagnosenet.github.io/
Participants: John Anderson Garcia Henao, Michel Riveill
Keywords: federated learning, low power deep learning
Collaborations: Felix Armendo Meija (Univ. Industrial Santander, Colombia) and Calos Jaime Barrios Hernandezt Bergé (Univ. Industrial Santander, Colombia)

DiagnoseNET (video on YouTube: https://www.youtube.com/watch?v=qM11jmYTmrs) [23] is designed as a modular framework that enables the deep learning application-workflow management and

Figure 2: The Topix platform for the co-clustering of text matrices.

expressivity to build and finetune the neural architecture, while its runtime abstracts the distributed orchestration of portability and scalability from a GPU workstation to multi-nodes computational platforms. It automatizes in one expression API the neural architecture definition, the hyperparameter search, the data locality and batching, while the runtime coordinate the parameters between the devices according to the execution modes, which enables the workers selection through synchronous or asynchronous coordination gradient computations with gRPC or MPI communication protocols, tested on x86 and arm architectures.

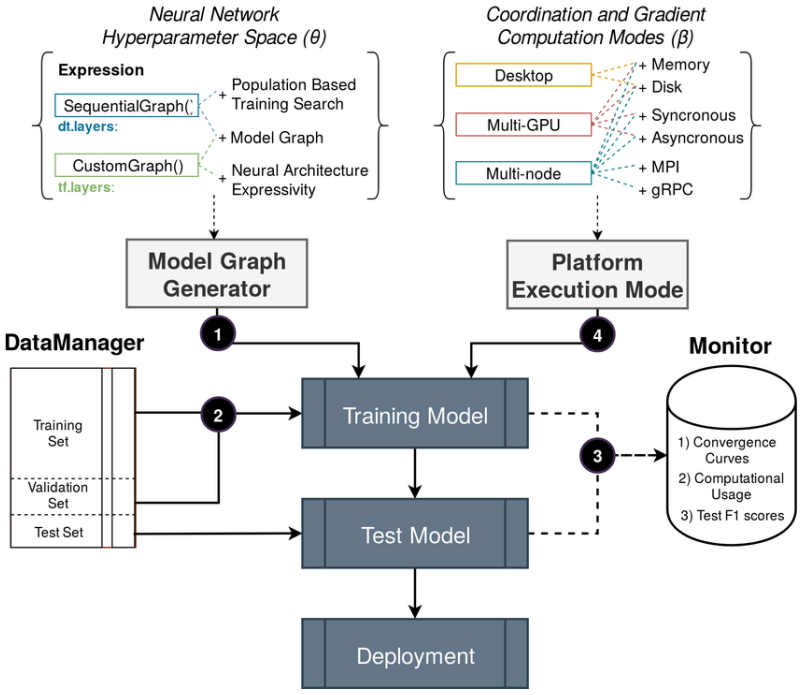


Figure 3: DiagnoseNet architecture.

7 New results

7.1 Unsupervised learning

7.1.1 Dynamic Co-Clustering

Participants: Charles Bouveyron, Marco Corneli, Giulia Marchello
Keywords: generative models, dynamic co-clustering, count data, pharmacovigilance
Collaborations: CHU de Nice (Centre de Pharmacovigilance)

We considered in [29, 30] the problem of co-clustering count matrices that may evolved along the time and we introduce a generative model to handle it. The proposed model, named dynamic latent block model, extend the classical latent block model to the dynamic case. The modeling of the dynamic in a continuous time relies on a non- homogeneous Poisson process, with a latent partition of time intervals. We proposed to use the SEM-Gibbs algorithm for model inference. An application to pharmacovigilance is currently under consideration.

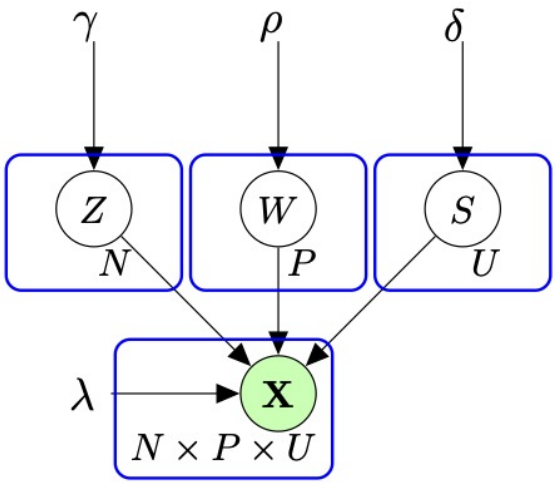


Figure 4: The generative model of the Dynamic Latent Block Model.

7.1.2 A Bayesian approach for clustering and exact finite-sample model selection in longitudinal data mixtures

Participants: Marco Corneli, Elena Erosheva
Keywords: longitudinal data, model based clustering, model selection, ICL, BIC, Gaussian processes

In [46], we consider mixtures of longitudinal trajectories, where one trajectory contains measurements over time of the variable of interest for one individual and each individual belongs to one cluster. The number of clusters as well as individual cluster memberships are unknown and must be inferred. We propose an original Bayesian clustering framework that allows us to obtain an exact finite-sample model selection criterion. Our approach is more flexible and parsimonious than asymptotic alternatives such as Bayesian Information Criterion (BIC) or Integrated Classification Likelihood (ICL) criterion in the choice of the number of clusters. Moreover, our approach has other desirable qualities: i) it keeps the computational effort of the clustering algorithm under control and ii) it generalizes to several families of regression mixture models, from linear to purely non-parametric.

7.1.3 Bayesian discriminative Gaussian clustering

Participants: Charles Bouveyron
Keywords: generative models, model-based clustering, Bayesian modeling, high-dimensional data,

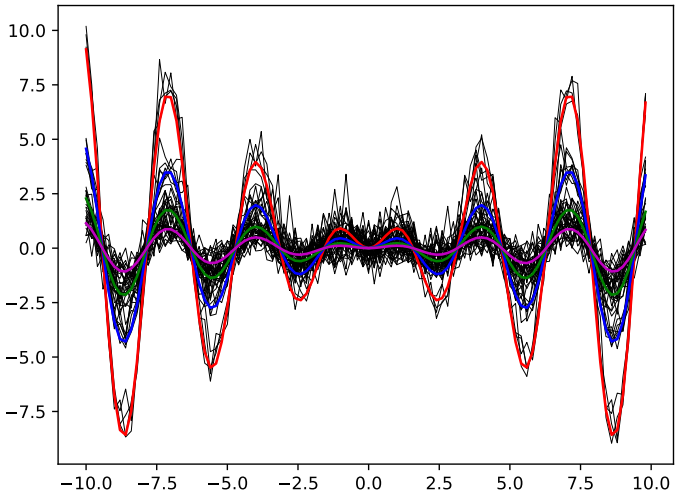


Figure 5: Bayesian fitting of non-parametric mixture regressions via the approach proposed in [46]

Collaborations: Pierre Latouche (Univ. de Paris), Nicolas Jouvin (Univ. Paris 1 & Institut Curie)

High-dimensional data clustering has become and remains a challenging task for modern statistics and machine learning, with a wide range of applications. We considered in [53] the powerful discriminative latent mixture model, and we extended it to the Bayesian framework. Modeling data as a mixture of Gaussians in a low-dimensional discriminative subspace, a Gaussian prior distribution is introduced over the latent group means and a family of twelve submodels are derived considering different covariance structures. Model inference is done with a variational EM algorithm, while the discriminative subspace is estimated via a Fisher-step maximizing an unsupervised Fisher criterion. An empirical Bayes procedure is proposed for the estimation of the prior hyper-parameters, and an integrated classification likelihood criterion is derived for selecting both the number of clusters and the submodel. The performances of the resulting Bayesian Fisher-EM algorithm are investigated in two thorough simulated scenarios, regarding both dimensionality as well as noise and assessing its superiority with respect to state-of-the-art Gaussian subspace clustering models. In addition to standard real data benchmarks, an application to single image denoising is proposed, displaying relevant results. This work comes with a reference implementation for the R software in the `FisherEM` package accompanying the paper.

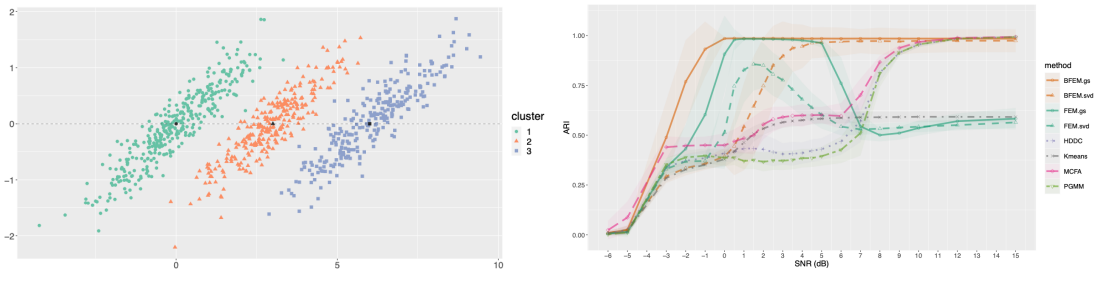


Figure 6: Comparison of the clustering performance of BFEM compared to competitive methods for (simulated) high-dimensional data. Numerical experiments showed that BFEM outperforms competitors in most situations, and in particular in difficult situations where the noise-to-signal ratio is very high.

7.1.4 A New Class of EM Algorithms. Escaping Local Maxima and Handling Intractable Sampling

Participants: Juliette Chevallier
Keywords: mixture models, variational inference
Collaborations: V. Debavelaere (CMAP), Stéphanie Allassonnière (Univ. de Paris)

The expectation-maximization (EM) algorithm is a powerful computational technique for maximum likelihood estimation in incomplete data models. When the expectation step cannot be performed in closed form, a stochastic approximation of EM (SAEM) can be used. The convergence of the SAEM toward critical points of the observed likelihood has been proved and its numerical efficiency has been demonstrated. However, sampling from the posterior distribution may be intractable or have a high computational cost. Moreover, despite appealing features, the limit position of this algorithm can strongly depend on its starting one. To cope with this two issues, we propose in [5] new stochastic approximation version of the EM in which we do not sample from the exact distribution in the expectation phase of the procedure. We first prove the convergence of this algorithm toward critical points of the observed likelihood. Then, we propose an instantiation of this general procedure to favor convergence toward global maxima. Experiments on synthetic and real data highlight the performance of this algorithm in comparison to the SAEM and the EM when feasible.

7.1.5 Co-Clustering of Multivariate Functional Data for Air Pollution Analysis

Participants: Charles Bouveyron
Keywords: generative models, model-based co-clustering, functional data, air pollution, public health
Collaborations: J. Jacques and A. Schmutz (Univ. de Lyon), Fanny Simoes and Silvia Bottini (MDlab, MIS, Univ. Côte d'Azur)

In [44], we focused on Air pollution, which is nowadays a major treat for public health, with clear links with many diseases, especially cardiovascular ones. The spatio-temporal study of pollution is of great interest for governments and local authorities when deciding for public alerts or new city policies against pollution raise. The aim of this work is to study spatio-temporal profiles of environmental data collected in the south of France (Région Sud) by the public agency AtmoSud. The idea is to better understand the exposition to pollutants of inhabitants on a large territory with important differences in term of geography and urbanism. The data gather the recording of daily measurements of five environmental variables, namely three pollutants (PM10, NO2, O3) and two meteorological factors (pressure and temperature) over six years. Those data can be seen as multivariate functional data: quantitative entities evolving along time, for which there is a growing need of methods to summarize and understand them. For this purpose, a novel co-clustering model for multivariate functional data is defined. The model is based on a functional latent block model which assumes for each co-cluster a probabilistic distribution for multivariate functional principal component scores. A Stochastic EM algorithm, embedding a Gibbs sampler, is proposed for model inference, as well as a model selection criteria for choosing the number of co-clusters. The application of the proposed co-clustering algorithm on environmental data of the Région Sud allowed to divide the region composed by 357 zones in six macro-areas with common exposure to pollution. We showed that pollution profiles vary accordingly to the seasons and the patterns are conserved during the 6 years studied. These results can be used by local authorities to develop specific programs to reduce pollution at the macro-area level and to identify specific periods of the year with high pollution peaks in order to set up specific prevention programs for health. Overall, the proposed co-clustering approach is a powerful resource to analyse multivariate functional data in order to identify intrinsic data structure and summarize variables profiles over long periods of time.

7.1.6 Semi-supervised Consensus Clustering Based on Frequent Closed Itemsets

Participants: Frédéric PRECIOSO (Université Côte d'Azur)
Keywords: Clustering; Semi-supervised learning; Semi-supervised consensus clustering; Frequent closed itemsets
Collaborations: Tianshu YANG (Université Côte d'Azur, Amadeus), Nicolas PASQUIER (Université Côte d'Azur), Antoine HOM (Amadeus), Laurent DOLLE (Amadeus), in a CIFRE PhD project with Amadeus

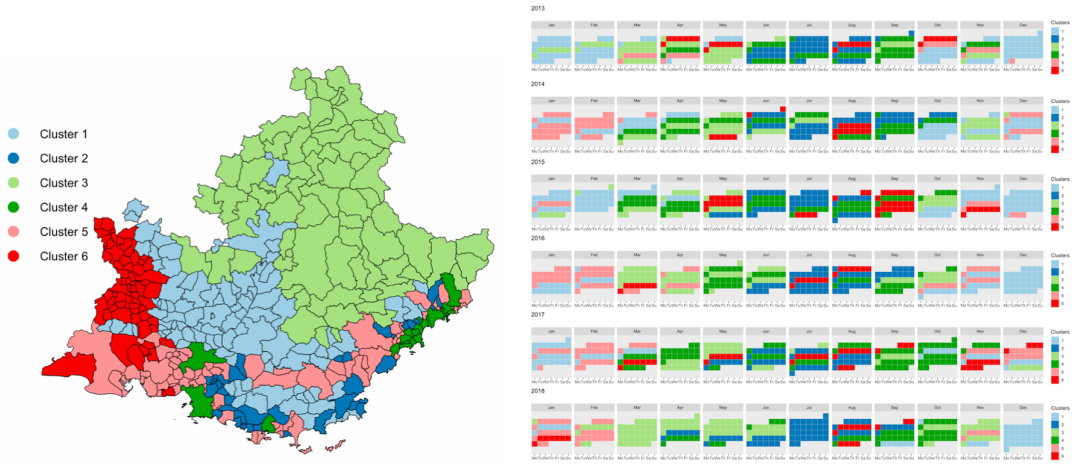


Figure 7: Spatial clustering of the area zones according to the air pollution dynamic over the studied period (left panel) and temporal segmentation of the time (right panel). Those tools may offer meaningful summaries on such massive pollution data to experts or local authorities.

Semi-supervised consensus clustering integrates supervised information into consensus clustering in order to improve the quality of clustering. In this paper, we study the novel Semi-MultiCons semi-supervised consensus clustering method extending the previous MultiCons approach [60]. Semi-MultiCons aims to improve the clustering result by integrating pairwise constraints in the consensus creation process and infer the number of clusters K using frequent closed itemsets extracted from the ensemble members. Experimental results show that the proposed method outperforms other state-of-art semi-supervised consensus algorithms.

7.1.7 Hierarchical clustering with discrete latent variable models and the ICL criterion

Participants: Charles Bouveyron
Keywords: generative models, model-based clustering, model selection, discrete latent variable models, networks
Collaborations: Pierre Latouche (Univ. de Paris), Nicolas Jouvin (Univ. Paris 1 & Institut Curie), E. Côme (Univ. Gustave Eiffel)

In [45], we introduce a two step methodology to extract a hierarchical clustering. This methodology considers the integrated classification likelihood criterion as an objective function, and applies to any discrete latent variable models (DLVM) where this quantity is tractable. The first step of the methodology involves maximizing the criterion with respect to the discrete latent variables state with uninformative priors. To that end we propose a new hybrid algorithm based on greedy local searches as well as a genetic algorithm which allows the joint inference of the number K of clusters and of the clusters themselves. The second step of the methodology is based on a bottom-up greedy procedure to extract a hierarchy of clusters from this natural partition. In a Bayesian context, this is achieved by considering the Dirichlet cluster proportion prior parameter α as a regularisation term controlling the granularity of the clustering. This second step allows the exploration of the clustering at coarser scales and the ordering of the clusters an important output for the visual representations of the clustering results. The clustering results obtained with the proposed approach, on simulated as well as real settings, are compared with existing strategies and are shown to be particularly relevant. This work is implemented in the R package `greed`.

7.1.8 Ensemble Clustering Based Semi-Supervised Learning for Revenue Accounting Workflow Management

Participants: Frédéric PRECIOSO (Université Côte d'Azur)

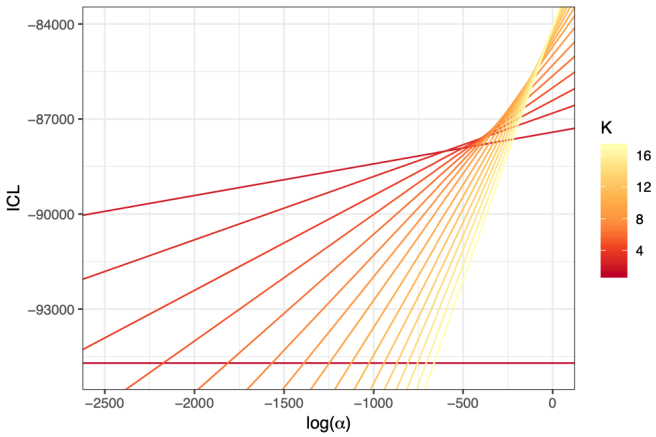


Figure 8: Lines of slope representing the ICL function according to log(α) for collections of partitions with a decreasing number of hierarchical clusters.

Keywords: Consensus Clustering; Closed sets; Anomalies correction; Multi-level clustering; Ensemble clustering; Revenue accounting workflow
Collaborations: Tianshu YANG (Université Côte d'Azur, Amadeus), Nicolas PASQUIER (Université Côte d'Azur), in a CIFRE PhD project with Amadeus

We present a semi-supervised ensemble clustering framework for identifying relevant multi-level clusters, regarding application objectives, in large datasets and mapping them to application classes for predicting the class of new instances [21]. This framework extends the MultiCons closed sets based multiple consensus clustering approach but can easily be adapted to other ensemble clustering approaches. It was developed to optimize the Amadeus S.A.S revenue accounting workflow management. Revenue accounting in travel industry is a complex task when travels include several transportations, with associated services, performed by distinct operators and on geographical areas with different taxes and currencies for example. Preliminary results show the relevance of the proposed approach for the automation of workflow anomaly corrections.

7.1.9 Clustering of count data through a mixture of multinomial PCA

Participants: Charles Bouveyron
Keywords: generative models, model-based clustering, count data, medicine
Collaborations: Pierre Latouche (Univ. de Paris), Nicolas Jouvin (Univ. Paris 1 & Institut Curie), Guillaume Bataillon and Alain Livartowski (Institut Curie)

In [10], we consider count data which are becoming more and more ubiquitous in a wide range of applications, with datasets growing both in size and in dimension. In this context, an increasing amount of work is dedicated to the construction of statistical models directly accounting for the discrete nature of the data. Moreover, it has been shown that integrating dimension reduction to clustering can drastically improve performance and stability. In this paper, we rely on the mixture of multinomial PCA, a mixture model for the clustering of count data, also known as the probabilistic clustering-projection model in the literature. Related to the latent Dirichlet allocation model, it offers the flexibility of topic modeling while being able to assign each observation to a unique cluster. We introduce a greedy clustering algorithm, where inference and clustering are jointly done by mixing a classification variational expectation maximization algorithm, with a branch and bound like strategy on a variational lower bound. An integrated classification likelihood criterion is derived for model selection, and a thorough study with numerical experiments is proposed to assess both the performance and robustness of the method. Finally, we illustrate the qualitative interest of the latter in a real-world application, for the clustering of anatomopathological medical reports, in partnership with expert practitioners from the Institut Curie hospital.

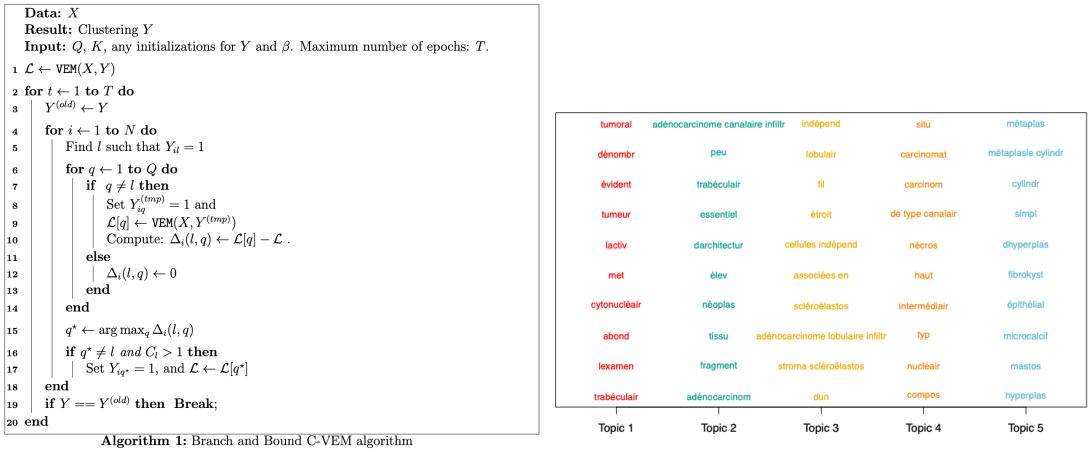


Figure 9: The mixture of multinomial PCA (MMPCA) algorithm (left panel) and its application for the clustering of medical reports (right panel).

7.1.10 Clustering multivariate functional data in group-specic functional subspaces

Participants: Charles Bouveyron
Keywords: generative models, model-based clustering, multivariate functional data, air pollution
Collaborations: J. Jacques and A. Schmutz (Univ. de Lyon)

With the emergence of numerical sensors in many aspects of every- day life, there is an increasing need in analyzing multivariate functional data. In [13], we focused on the clustering of such functional data, in order to ease their modeling and understanding. To this end, a novel clustering technique for multivariate functional data is presented. This method is based on a functional latent mixture model which fits the data in group-specific functional subspaces through a multivariate functional principal component analysis. A family of parsimonious models is obtained by constraining model parameters within and between groups. An EM algorithm is proposed for model inference and the choice of hyper-parameters is addressed through model selection. Numerical experiments on simulated datasets highlight the good performance of the proposed methodology compared to existing works. This algorithm is then applied to the analysis of the pollution in French cities for one year.

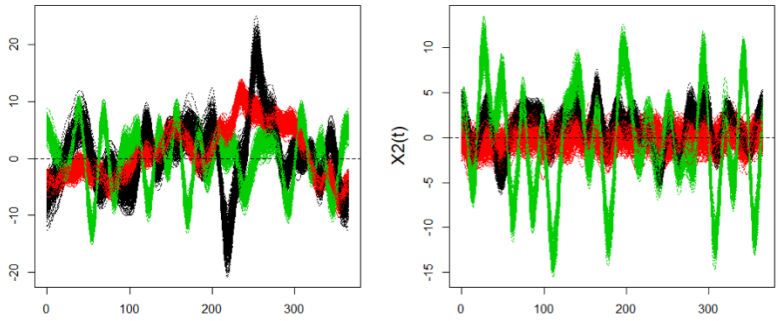


Figure 10: The funHDDC algorithm allows to cluster multivariate functional data. Here, we applied it to multivariate time series of air polluants in France.

7.1.11 Exact Bayesian model selection for principal component analysis

Participants: Charles Bouveyron, Pierre-Alexandre Mattei

Keywords: Model selection, generative models, Bayesian modeling, high-dimensional data
Collaborations: Pierre Latouche (Univ. de Paris)

Principal component analysis (PCA) is the most commonly used method to reduce the dimensionality of a data set. By linearly transforming p features into d principal components, PCA somehow allows to bypass the curse of dimensionality. A key question pertaining the use of PCA is *what number d of principal components should we choose?* The framework of Bayesian model selection, that we recently reviewed in [57], provides, in theory, a satisfying answer. Indeed, by relying on a probabilisitic model for PCA, it allows to find the dimensionality with largest posterior probability. However, there are two main practical obstacles that prevent to perform Bayesian model selection for PCA: *(i)* the fact that marginal likelihood of Bayesian PCA models are very difficult to compute, and *(ii)* the question of the choice of a relevant prior distribution. In this work [6], we address both issues and propose a simple way to perform exact model selection for Bayesian PCA. To this end, we introduce a novel Bayesian formulation of PCA based on a normal-gamma prior distribution. In this context, we exhibit a closed-form expression of the marginal likelihood which allows to infer an optimal number of components. This allows to address *(i)* in a simple way: we can compute the marginal likelihood exactly, rather than having to rely on approximations or expensive Monte Carlo sampling. To address *(ii)*, we propose a heuristic based on the expected shape of the marginal likelihood curve in order to choose the hyperparameters. In nonasymptotic frameworks, we show on simulated data that this exact dimensionality selection approach is competitive with both Bayesian and frequentist state-of-the-art methods.

7.1.12 International Patent: Clustering Techniques for Revenue Accounting Error-Handling Automation

Participants: Frédéric PRECIOSO (Université Côte d'Azur)
Keywords: Revenue accounting workflow; Semi-supervised consensus clustering; Frequent closed itemsets;
Collaborations: Tianshu YANG (Université Côte d'Azur, Amadeus), Nicolas PASQUIER (Université Côte d'Azur), Antoine HOM (Amadeus), Laurent DOLLE (Amadeus) in a CIFRE PhD project with Amadeus

This invention proposed by Amadeus SAS and Université Cote d'Azur (UCA) concerns the application of machine learning techniques for the revenue accounting task management [61]. The revenue accounting system automatically process documents until an error occurs. The process is then interrupted and set to manual for validation or error correction, resulting in additional costs and delays. This invention aims to decrease cost time and manual efforts that Amadeus and its customers spend on tasks [20]. This goal will be achieved by: 1) automatic identification of anomaly patterns through the unsupervised clustering of error tickets to form clusters of tickets corresponding to similar anomalies and requiring similar correction processes 2) Automatic or semi-automatic, depending on the type of the anomaly pattern, validation and/or correction of the error by the analysis of logs of correction actions taken by the correctors, associated to each cluster of tickets for the automation of the error correction process.

7.2 Understanding (deep) learning models

7.2.1 Theoretical properties of Importance Weighted variational inference

Participants: Pierre-Alexandre Mattei
Keywords: deep learning, neural networks, generative models, variational inference
Collaborations: Jes Frellsen (Technical University of Denmark)

Importance weighted variational inference (IWVI) is a promising strategy for learning latent variable models. IWVI uses new variational bounds, known as Monte Carlo objectives (MCOs), obtained by replacing intractable integrals by Monte Carlo estimates—usually simply obtained via importance sampling. These bounds $\mathscr{L}_K(Q)$ essentially depend on two things: the number of samples used K, and the distribution of the importance weights Q. It was previously shown that increasing the number of importance samples L provably tightens the gap between the bound and the likelihood. Inspired by this simple monotonicity theorem, we present a series of nonasymptotic results that link properties of Monte Carlo estimates to tightness of MCOs [31]. We challenge in particular the rationale that smaller Monte

Carlo variance leads to better bounds. Moreover, we show that increasing the negative dependence of importance weights monotonically increases the bound. To this end, we use the supermodular order $\preceq$ as a measure of dependence. Roughly speaking, for two random vectors $\mathbf{w} \sim Q_1$ and $\mathbf{v} \sim Q_2$, $Q_1 \preceq Q_2$ means that the coordinates of $\mathbf{w} \sim Q_1$ are more negatively dependent that those of $\mathbf{v} \sim Q_2$). We then show that negative dependence tightens the bound in the following sense : $Q_1 \preceq Q_2$ implies $\mathscr{L}_K(Q_1) \geq \mathscr{L}_K(Q_2)$ Our simple result provides theoretical support to several different approaches that leveraged negative dependence to perform efficient variational inference of deep generative models.

7.2.2 T-CAM: class activation maps for text classification

Participants: Frederic Precioso, Marco Corneli, Laurent Vanni
Keywords: Text classification, Convolutional Neural Networks, Recurrent Neural Networks, Class activation map
Collaborations: Damon Mayaffre, CNRS

The main aim of the works [43, 40] is to bring interpretability of deep networks for text classification to another level. Although some popular feature extraction techniques (e.g.attention in RNNs or Text Deconvolution Saliency for CNNs) can detect relevant linguistic markers used by a neural net to achieve classification, the same techniques are unable to assess how these markers are used. In other words, they cannot indicate whether a relevant linguistic marker provides support for or against aclass assignment. In order to overcome this limitation, we generalize the Class Activation Map (CAM) technique from computer vision to textual feature extraction. The new text CAM (T-CAM) simultaneously assesses the positive/negative contribution of any relevant input feature with respect to any class and it can be computed for several architectures (including CNNs and RNNs). Finally, T-CAM requires no sampling, neither binarization in multiclass classification problems, thus making it an attractive alternative to LIME. This work has led to several other contributions in the field of linguistics [33, 32].

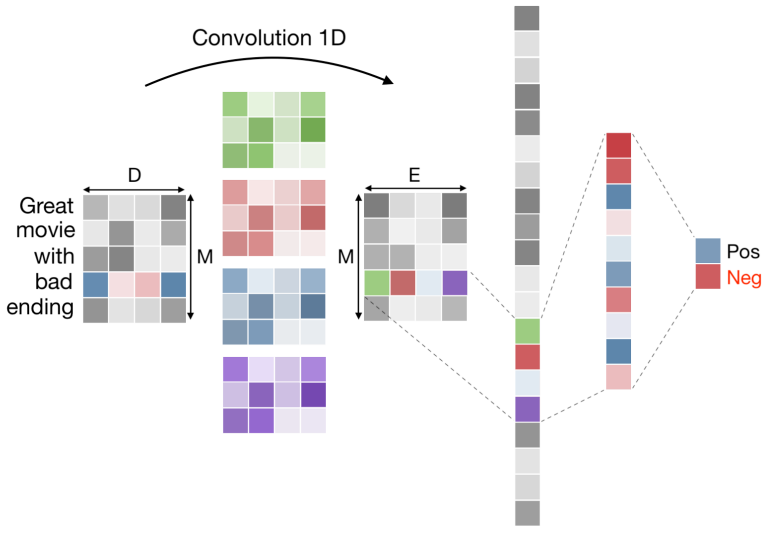


Figure 11: T-CAM applied to CNNs

7.2.3 Explaining the explainer: a first theoretical analysis of LIME

Participants: Damien Garreau
Keywords: machine learning, interpretability, statistical learning theory
Collaborations: Ulrike von Luxburg (Tübingen University)

Machine learning is used more and more often for sensitive applications, sometimes replacing humans in critical decision-making processes. As such, interpretability of these algorithms is a pressing

need. One popular algorithm to provide interpretability is LIME (Local Interpretable Model-Agnostic Explanation). In [24], we provide the first theoretical analysis of LIME. We derive closed-form expressions for the coefficients of the interpretable model when the function to explain is linear. The good news is that these coefficients are proportional to the gradient of the function to explain: LIME indeed discovers meaningful features. However, our analysis also reveals that poor choices of parameters can lead LIME to miss important features.

7.2.4 Looking deeper into tabular LIME

Participants: Damien Garreau
Keywords: machine learning, interpretability, statistical learning theory
Collaborations: Ulrike von Luxburg (Tübingen University)

Interpretability of machine learning algorithm is a pressing need. Numerous methods appeared in recent years, but do they make sense in simple cases? In [50], we present a thorough theoretical analysis of Tabular LIME. In particular, we show that the explanations provided by Tabular LIME are close to an explicit expression in the large sample limit. We leverage this knowledge when the function to explain has some nice algebraic structure (linear, multiplicative, or depending on a subset of the coordinates) and provide some interesting insights on the explanations provided in these cases. In particular, we show that Tabular LIME provides explanations that are proportional to the coefficients of the function to explain in the linear case, and provably discards coordinates unused by the function to explain in the general case.

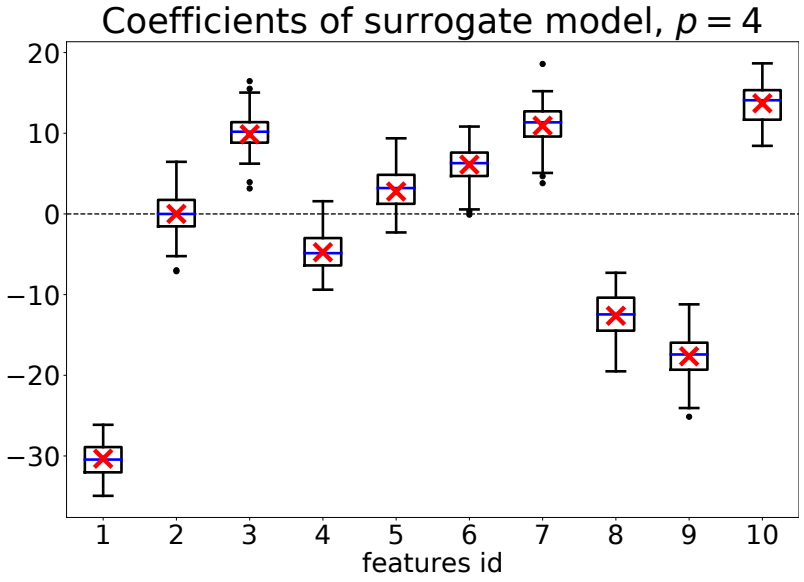


Figure 12: Theoretical predictions *vs* practice for tabular LIME.

7.2.5 An analysis of LIME for text data

Participants: Damien Garreau
Keywords: machine learning, interpretability, statistical learning theory, natural language processing
Collaborations: Dina Mardaoui (Polytech Nice)

Text data are increasingly handled in an automated fashion by machine learning algorithms. But the models handling these data are not always well-understood due to their complexity and are more and more often referred to as "black-boxes." Interpretability methods aim to explain how these models operate. Among them, LIME has become one of the most popular in recent years. However, it comes without theoretical guarantees: even for simple models, we are not sure that LIME behaves accurately. In [56], we provide a first theoretical analysis of LIME for text data. As a consequence of our theoretical findings, we

show that LIME indeed provides meaningful explanations for simple models, namely decision trees and linear models.

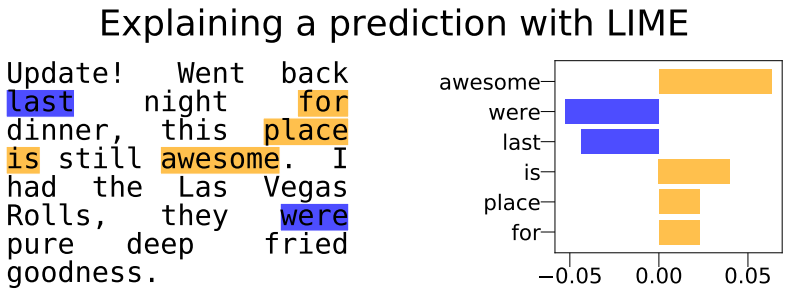


Figure 13: Explaining a prediction with LIME

7.2.6 Visualizing ECG Contribution into Convolutional Neural Network Classification

Participants: Frédéric Precioso
Keywords: machine learning, interpretability, statistical learning theory, natural language processing
Collaborations: Yaowei Li (UCA & School of Instrument Science and Engineering, Southeast University Nanjing, China), Chengyu Liu (The State Key Laboratory of Bioelectronics, School of Instrument Science and Engineering Southeast University Nanjing, China)

Convolutional Neural Network (CNN) demonstrated impressive classification performance on ElectroCardioGram (ECG) analysis and has a great potential to extract salient patterns from signal. Visualizing local contributions of ECG input to a CNN classification model can both help us understand why CNN can reach such high-level performances and serve as a basis for new diagnosis recommendations. In [42], we build a single-layer 1-D CNN model on ECG classification with a 99% accuracy [42]. The trained model is then used to build a 1-D Deconvolved Neural Network (1-D DeconvNet) for visualization. We propose Feature Importance Degree Heatmap (FIDH) to interpret the contribution of each point of ECG input to CNN classification results, and thus to show which part of ECG raises attention of the classifier. We also illustrate the correlation between two visualization methods: first-order Taylor expansion and multilayer 1-D DeconvNet.

7.3 Adaptive and robust learning

7.3.1 Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification

Participants: Frédéric Precioso
Keywords: machine learning, cumulative learning, transfer learning, medical domain
Collaborations: Khawla Seddiki (Univ. Laval, Quebec, CA), Philippe Saudemont (Université de Lille, Inserm), Nina Ogrinc (Université de Lille, Inserm), Maxence Wisztorski (Université de Lille, Inserm), Michel Salzet (Université de Lille, Inserm), Isabelle Fournier (Université de Lille, Inserm), Arnaud Droit (Univ. Laval, Quebec, CA)

Rapid and accurate clinical diagnosis remains challenging. A component of diagnosis tool development is the design of effective classification models with Mass spectrometry (MS) data. Some Machine Learning approaches have been investigated but these models require time-consuming preprocessing steps to remove artifacts, making them unsuitable for rapid analysis. Convolutional Neural Networks (CNNs) have been found to perform well under such circumstances since they can learn representations from raw data. However, their effectiveness decreases when the number of available training samples is small, which is a common situation in medicine. In this work, we investigate transfer learning on 1D-CNNs, then we develop a cumulative learning method when transfer learning is not powerful enough. We propose to train the same model through several classification tasks over various small datasets to

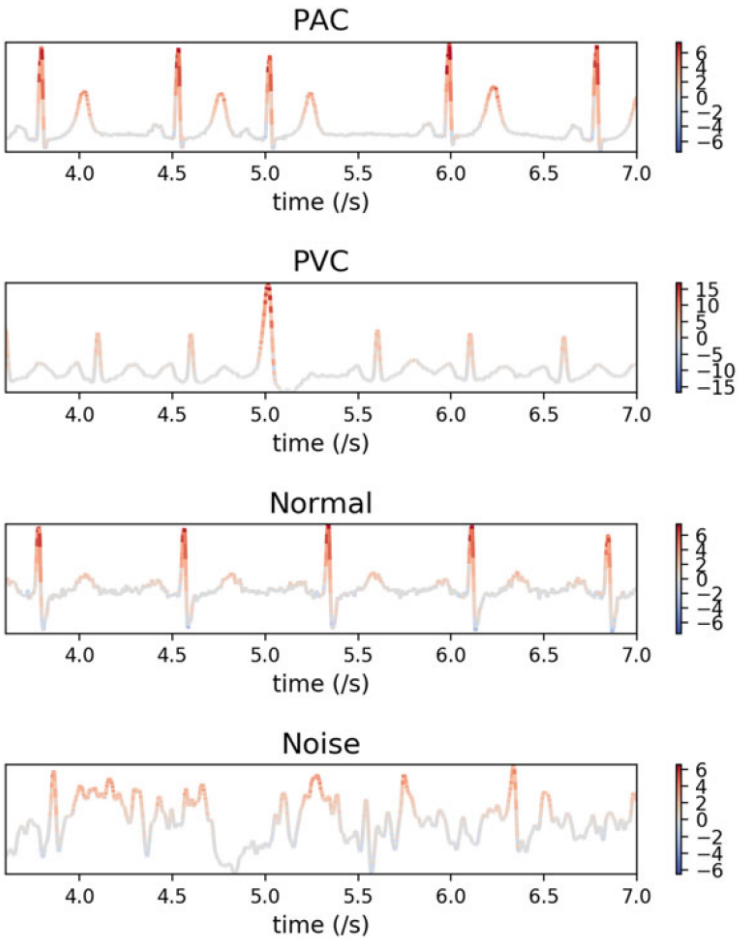


Figure 14: Feature Importance Degree Heatmap (FIDH) of four types of signal: PAC (Premature Atrial Contraction), PVC (Premature Ventricular Contraction), Normal beat, and Noise beat

accumulate knowledge in the resulting representation [14]. By using rat brain as the initial training dataset, a cumulative learning approach can have a classification accuracy exceeding 98% for 1D clinical MS-data. We show the use of cumulative learning using datasets generated in different biological contexts, on different organisms, and acquired by different instruments. Here we show a promising strategy for improving MS data classification accuracy when only small numbers of samples are available.

7.3.2 NEWMA: a new method for scalable model-free online change-point detection

Participants: Damien Garreau
Keywords: change-point detection, time series, kernel methods
Collaborations: Nicolas Keriven (GIPSA lab), Iacopo Poli (LightOn)

We consider the problem of detecting abrupt changes in the distribution of a multi-dimensional time series, with limited computing power and memory. In [11], we propose a new, simple method for model-free online change-point detection that relies only on fast and light recursive statistics, inspired by the classical Exponential Weighted Moving Average algorithm (EWMA). The proposed idea is to compute two EWMA statistics on the stream of data with different forgetting factors, and to compare them. By doing so, we show that we implicitly compare recent samples with older ones, without the need to explicitly store them. Additionally, we leverage Random Features (RFs) to efficiently use the Maximum Mean Discrepancy as a distance between distributions, furthermore exploiting recent optical hardware

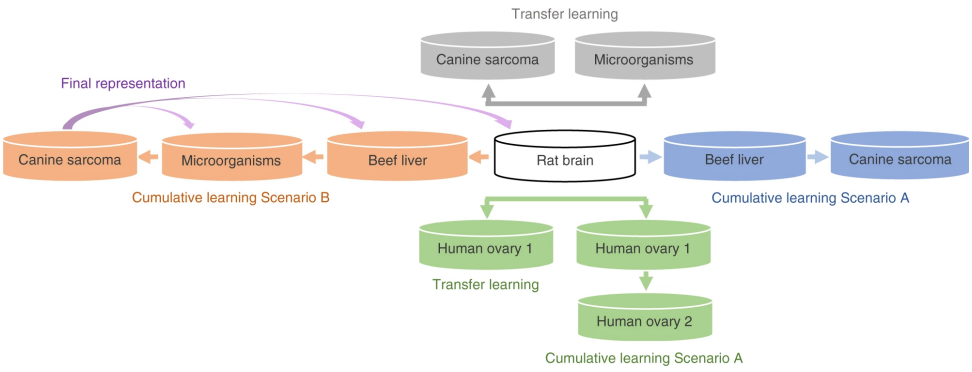


Figure 15: By cumulative learning Scenario A (in blue for canine sarcoma and in green for human ovary 2). By cumulative learning Scenario B (in orange for canine sarcoma). Final representation of Scenario B is tested on the datasets used during the training (in purple arrows).

to compute high-dimensional RFs in near constant time. We show that our method is significantly faster than usual non-parametric methods for a given accuracy.

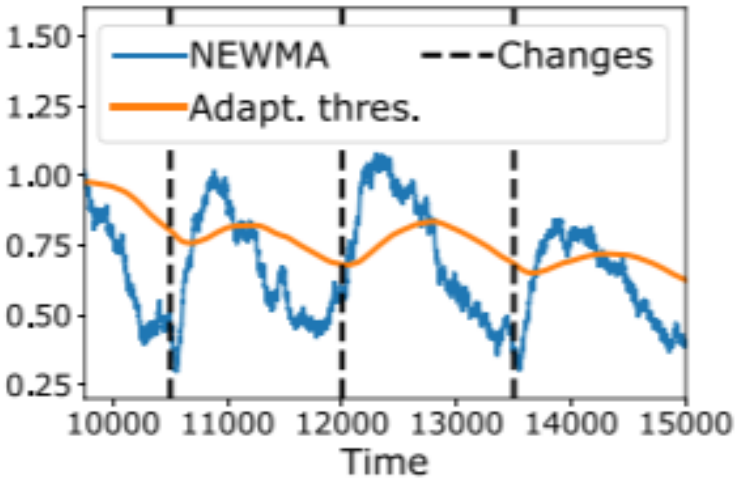


Figure 16: NEWMA.

7.4 Learning with heterogeneous and corrupted data

7.4.1 Co-Clustering of ordinal data via latent continuous random variables and Not Missing at Random Entries

Participants: Charles Bouveyron, Marco Corneli
Keywords: generative models, model-based clustering, count data, medicine
Collaborations: Pierre Latouche (Univ. de Paris)

This work [8] is about the co-clustering of ordinal data. Such data are very common on e-commerce platforms where customers rank the products/services they bought. In more detail, we focus on arrays of ordinal (possibly missing) data involving two disjoint sets of individuals/objects corresponding to the

rows/columns of the arrays. Typically, an observed entry (i, j) in the array is an ordinal score assigned by the individual/row i to the object/column j. A new generative model for arrays of ordinal data is introduced along with an inference algorithm for parameters estimation. The model accounts for not missing at random data and relies on latent continuous random variables. The fitting allows to simultaneously co-cluster the rows and columns of an array. The estimation of the model parameters is performed via a classification expectation maximization algorithm. A model selection criterion is formally obtained to select the number of row and column clusters. To show that our approach reaches and often outperforms the state of the art, we carry out numerical experiments on synthetic data. Finally, applications on real datasets highlight the model capacity to deal with very sparse arrays.

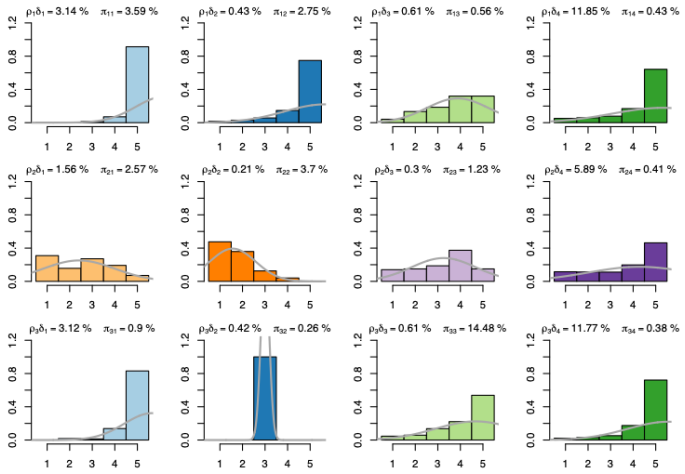


Figure 17: Result of the clustering of ordinal data of Amazon fine foods with the oLBM algorithm we proposed. The output of the method allows a clear understanding of the latent continuous variable that users supposedly use for grading products.

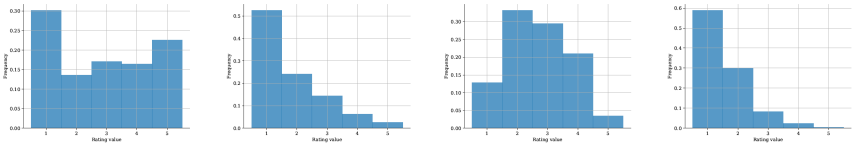
7.4.2 Deep generative modelling with missing not at random data

Participants: Pierre-Alexandre Mattei
Keywords: missing data, neural networks, deep learning, generative models,
Collaborations: Jes Frellsen, Niel Bruun Ipsen (Technical University of Denmark)

When a missing process depends on the missing values themselves, it needs to be explicitly modelled and taken into account while doing likelihood-based inference. We present an approach for building and fitting deep latent variable models (DLVMs) in cases where the missing process is dependent on the missing data. Specifically, a deep neural network enables us to flexibly model the conditional distribution of the missingness pattern given the data. This allows for incorporating prior information about the type of missingness (e.g. self-censoring) into the model. Our inference technique, based on importance-weighted variational inference, involves maximising a lower bound of the joint likelihood. Stochastic gradients of the bound are obtained by using the reparameterisation trick both in latent space and data space. Our method is called *not-missing-at-random importance-weighted autoencoder (not-MIWAE)* [17]. We show on various kinds of data sets and missingness patterns that explicitly modelling the missing process can be invaluable. We apply our method to censoring in datasets from the UCI database, clipping in images and the issue of selection bias in recommender systems.

7.4.3 DeepLTRS: A Deep Latent Recommender System based on User Ratings and Reviews

Participants: Dingge Liang, Marco Corneli, Charles Bouveyron.
Keywords: Recommender System, Topic modelling, latent representation learning.
Collaborations: Pierre Latouche (Univ. de Paris)



(a) MNAR train samples (b) MCAR test samples (c) MIWAE impute (d) not-MIWAE impute

Figure 18: Using not-MIWAE to tackle selection bias in a recommender system. Yahoo! Histograms over rating values from (a) the MNAR training set and (b) the MCAR test set. (c) and (d) show histograms over imputations of missing values in the test set, when encoding the corresponding training set. The not-MIWAE imputations (d) are much more faithful to the shape of the test set (b) than the MIWAE imputations (c).

We introduce in [28, 54] a deep latent recommender system (deepLTRS) in order to provide users with high quality recommendations based on observed user ratings and texts of product reviews. Our approach adopts a variational auto-encoder architecture as a generative deep latent variable model for both an ordinal matrix encoding users scores about products, and a document-term matrix encoding the reviews. Moreover, an alternated user/product mini-batching optimization structure is proposed to jointly capture user and product preferences. Numerical experiments on simulated and real-world data sets demonstrate that deepLTRS outperforms the state-of-the-art, in particular in contexts of extreme data sparsity.

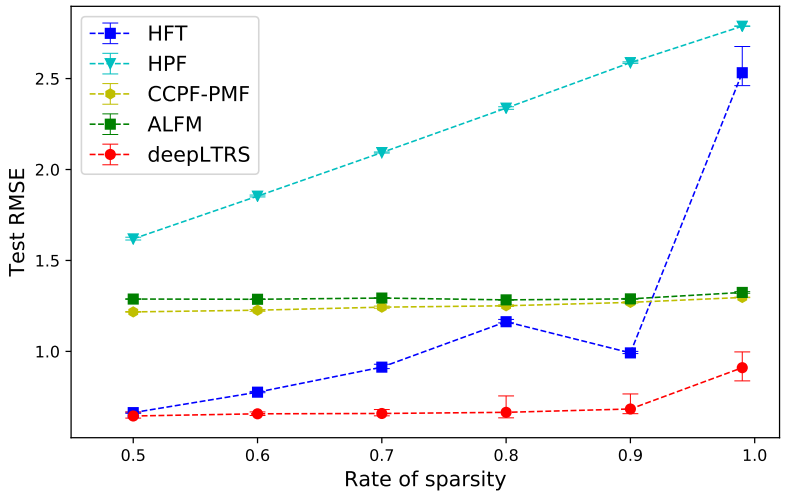


Figure 19: Test RMSE of models with different sparsity level on simulated data.

7.4.4 Hierarchical Multimodal Attention for Deep Video Summarization

Participants: Melissa Sanabria, Frédéric Precioso, (Inria, CNRS, I3S, Maasai, Université Côte d'Azur).
Keywords: Event stream data, Soccer match data, Video Summarization, Multimodal data, Sports Analytics
Collaborations: Thomas Menguy (Wildmoka Company), this work has been co-funded by Région Sud Provence Alpes Côte d'Azur (PACA), Université Côte d'Azur (UCA) and Wildmoka Company.

This paper explores the problem of summarizing professional soccer matches as automatically as possible using both the event-stream data collected from the field and the content broadcasted on TV. We have designed an architecture, introducing first (1) a Multiple Instance Learning method that takes into account the sequential dependency among events and then (2) a hierarchical multimodal attention

layer that grasps the importance of each event in an action [38]. We evaluate our approach on matches from two professional European soccer leagues, showing its capability to identify the best actions for automatic summarization by comparing with real summaries made by human operators.

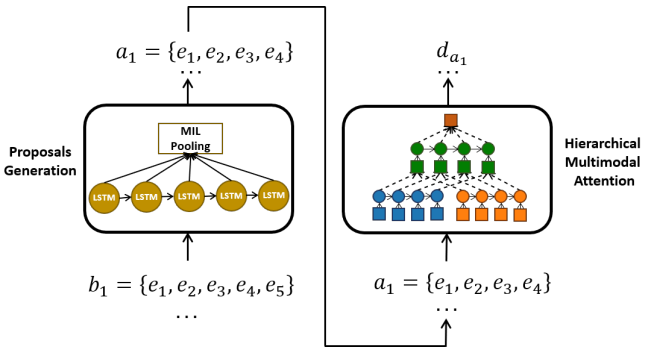


Figure 20: General schema of our approach. The left part of the figure represents the first block of our approach: Proposals Generation with a LSTM MIL Pooling. It gets as input the bags of events and outputs action Proposals. The right part of the figure is the second block of our approach: Hierarchical Multimodal Attention. It gets as input the action proposals (events data and audio data) and predict the likelihood for the given action to be in the summary.

7.4.5 Profiling Actions for Sport Video Summarization: An attention signal analysis

Participants: Melissa Sanabria, Frédéric Precioso, (Inria, CNRS, I3S, Maasai, Université Côte d'Azur).
Keywords: Video Summarization, Sports Video, Event stream data, Neural Network, User-Centric
Collaborations: Thomas Menguy (Wildmoka Company), this work has been co-funded by Région Sud Provence Alpes Côte d'Azur (PACA), Université Côte d'Azur (UCA) and Wildmoka Company.

Currently, in broadcast companies many human operators select which actions should belong to the summary based on multiple rules they have built upon their own experience using different sources of information. These rules define the different profiles of actions of interest that help the operator to generate better customized summaries. Most of these profiles do not directly rely on broadcast video content but rather exploit metadata describing the course of the match. In [19], we show how the signals produced by the attention layer of a recurrent neural network can be seen as a learned representation of these action profiles and provide a new tool to support operators' work. The results in soccer matches show the capacity of our approach to transfer knowledge between datasets from different broadcasting companies, from different leagues, and the ability of the attention layer to learn meaningful action profiles.

7.4.6 A coherent framework for learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data

Participants: Juliette Chevallier
Keywords: Bayesian inference, differential geometry, manifold-valued data
Collaborations: V. Debavelaere (CMAP), Stéphanie Allassonnière (Univ. de Paris)

We provide in [7] a coherent framework for studying longitudinal manifold-valued data. We introduce a Bayesian mixed-effects model which allows estimating both a group-representative piecewise-geodesic trajectory in the Riemannian space of shape and inter-individual variability. We prove the existence of the maximum a posteriori estimate and its asymptotic consistency under reasonable assumptions. Due to the non-linearity of the proposed model, we use a stochastic version of the Expectation-Maximization algorithm to estimate the model parameters. Our simulations show that our model is not noise-sensitive and succeeds in explaining various paths of progression.

7.4.7 How to deal with missing data in supervised deep learning?

Participants: Pierre-Alexandre Mattei
Keywords: missing data, neural networks, deep learning, generative models,
Collaborations: Jes Frellsen, Niel Bruun Ipsen (Technical University of Denmark)

The issue of missing data in supervised learning has been largely overlooked, especially in the deep learning community. In [26], we investigate strategies to adapt neural architectures to handle missing values. Here, we focus on regression and classification problems where the features are assumed to be missing at random. Of particular interest are schemes that allow to reuse as-is a neural discriminative architecture. One scheme involves imputing the missing values with learnable constants. We propose a second novel approach that leverages recent advances in deep generative modelling. More precisely, a deep latent variable model can be learned jointly with the discriminative model, using importance-weighted variational inference in an end-to-end way. This hybrid approach, which mimics multiple imputation, also allows to impute the data, by relying on both the discriminative and the generative model. We also discuss ways of using a pre-trained generative model to train the discriminative one. In domains where powerful deep generative models are available, the hybrid approach leads to large performance gains.

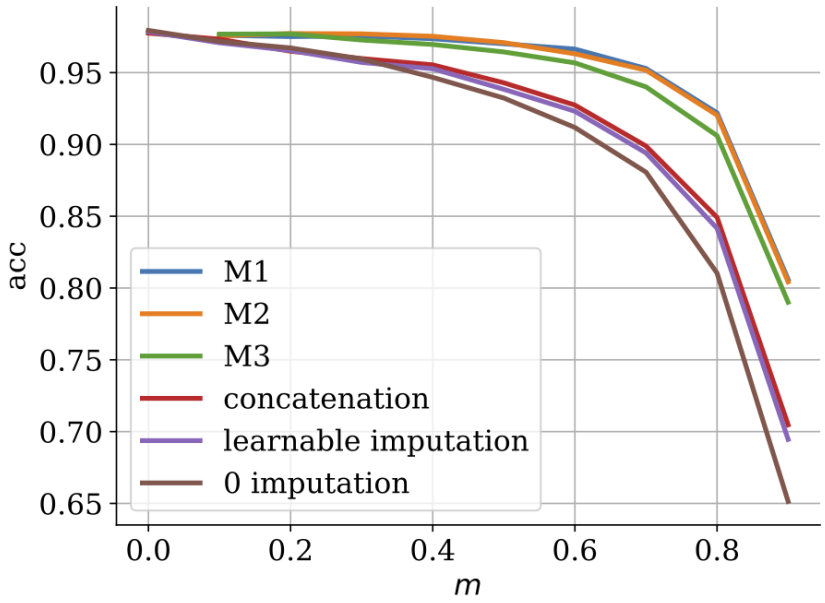


Figure 21: Performance of several strategies that we propose to deal with missing values (M1, M2, M3) on the MNIST data set, compared with other baselines.

7.4.8 NIH Peer Review: Criterion Scores Completely Account for Racial Disparities in Overall Impact Scores

Participants: Elena A. Erosheva
Keywords: grant funding, hierarchical models, matching, multilevel models, racial disparities
Collaborations: Sheridan Grant, Mei-Ching Chen, Mark D. Lindner, Richard K. Nakamura, Carole J. Lee

Previous research has found that (a) funding disparities are driven by applications' final impact scores and (b) only a portion of the black/white funding gap can be explained by bibliometrics and topic choice. Using NIH R01 applications for council years 2014-2016, we show in [9] that black/white funding disparities remain. We then examine the relationship between assigned reviewers' preliminary overall impact scores and preliminary criterion scores—Significance, Investigator, Innovation, Approach, and Environment—to evaluate whether racial disparities in preliminary overall impact scores can be explained

by other application and applicant characteristics. We hypothesize that differences in commensuration—the process of combining criterion scores into overall impact scores—disadvantage black applicants. Using multilevel models and matching on key variables including career stage, gender, and area of science to reduce model dependence, we find little evidence for racial disparities emerging in the process of combining preliminary criterion scores into preliminary overall impact scores. At the same time, we find that preliminary criterion scores fully account for racial disparities—yet do not explain all of the variability—in preliminary overall impact scores. Future research should focus on understanding the extent to which disparities in preliminary criterion scores could be driven by implicit racial preferences, black-white differences in bibliometric outputs, topic choice, and/or current or cumulative experiences impacting grantsmanship.

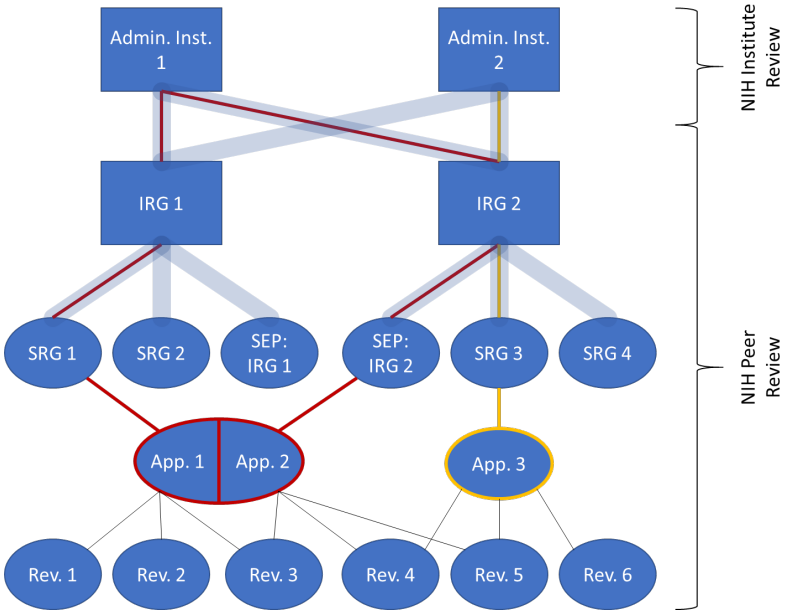


Figure 22: Multilevel NIH review structure for a hypothetical example of three applications (App. 1, 2, and 3) submitted by two PIs (yellow and red). Thick blue lines show structural connections. Thin lines show hypothetical assignments for the three applications. Rectangles are specified as fixed effects and ellipses as random effects in our mixed effects models.

7.4.9 Machine Learning Approaches to Epidemic Models

Participants: Marco Corneli, Marco Gori
Keywords: Epidemic models, spatiotemporal diffusion models, variational learning models
Collaborations: Andrea Zugarini (UNISI), Enrico Meloni (UNISI), Alessandro Betti (UNISI)

The COVID-19 outbreak has stimulated the interest in the proposal of novel epidemiological models to predict the course of the epidemic so as to help planning effective control strategies. In particular, in order to properly interpret the available data, it has become clear that one must go beyond most classic epidemiological models and consider richer description of the stages of infection. The problem of learning the parameters of these models is of crucial importance especially when assuming that they are time-variant, which further enriches their effectiveness. In this project [59] we propose a general approach for learning time-variant parameters of dynamic compartmental models from epidemic data. We formulate the problem in terms of a functional risk that depends on the learning variables through the solutions of a dynamic system. The resulting variational problem is then solved by using a gradient flow on a suitable, regularized functional. We forecast the epidemic evolution in Italy and France.

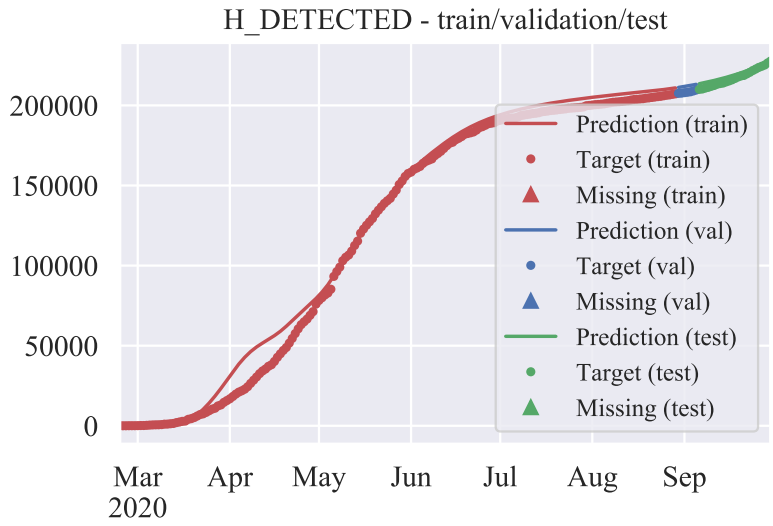


Figure 23: Fitting and prediction of the (detected) COVID-19 healed individuals in Italy according to the SIDARTHE Epidemic model.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

8.1.1 Orange

Participants: Hugo Miralles, Michel Riveill
External participants: Tamara Tosic (Orange), Thierry Nagellen (Orange)
Value: 45000 EUR

This collaboration contract is a CIFRE contract built upon the PhD of Hugo Miralles on Distributed device-embedded classification and prediction in near-to-real time. In this thesis we study the problem of efficient classification and prediction of multivariate time-series captured by embedded devices by using joint data-model distributed algorithms for applications that preserve private data.

8.1.2 Pro-BTP

Participants: Michel Riveill, Mansour Zoubeirou
External participants: Frédéric Estroumza (Pro BTP), Philippe Darnault (Pro BTP)
Value: Under negotiation

This collaboration contract is a CIFRE contract built upon the PhD of Mansour Zoubeirou on Detection of weak signals and incremental evolution of prediction models, especially in the context of fraud detection.

8.1.3 NXP

Participants: Frederic Precioso, Charles Bouveyron, Baptiste Pouthier (Ph.D. candidate)
External participants: Laurent Pilati (NXP)
Value: 45000 EUR

This collaboration contract is a CIFRE contract built upon the PhD of Baptiste Pouthier on Deep Learning and Statistical Learning on audio-visual data for embedded systems.

8.1.4 Oscaro

Participants: Charles Bouveyron, Pierre-Alexandre Mattei

External participants: Warith Harchaoui (Ph.D. candidate), Nils Grunwald (Oscaro)
Value: 45000 EUR

This collaboration contract is a CIFRE contract built upon the PhD of Warith Harchaoui on representation learning using neural networks and optimal transport.

8.1.5 Ezako

Participant: Frederic Precioso
External participants: Mireille Blay-Fornarino (Univ. Côte d'Azur), Yassine El Amraoui (Ezako - Univ. Côte d'Azur), Julien Muller (Ezako), Bora Kizil (Ezako)
Value: 45000 EUR

This collaboration contract is a CIFRE contract built upon the PhD of Yassine El Amraoui on Maximizing expert feedback in the detection of anomalies in time series.

8.1.6 Amadeus

Participant: Frederic Precioso
External participants: Nicolas Pasquier (Univ. Côte d'Azur), Tianshu Yang (Amadeus - Univ. Côte d'Azur), Antoine Hom (Amadeus), Laurent Dolle (Amadeus), Mickael Defoin-Platel (Amadeus), Luca Marchetti (Amadeus)
Value: 60000 EUR

This collaboration contract is a CIFRE contract built upon the PhD of Tianshu Yang on Semi-supervised clustering applied in revenue accounting.

8.1.7 Detection and characterization of salient moments for automatic summaries

Participants: Melissa Sanabria, Frédéric Precioso.
External Participants: Thomas Menguy (Wildmoka)
Value: 45000 EUR
Keywords: Video Summarization, Multimodal data, Soccer match data.

We have designed an architecture, introducing a Multiple Instance Learning method that takes into account the sequential dependency among events and a hierarchical multimodal attention layer that grasps the importance for each event in an action. We evaluate our approach on matches from two professional European soccer leagues, showing its capability to identify the best actions for automatic summarization by comparing with real summaries made by human operators.

8.2 Bilateral grants with industry

8.2.1 Grant from the Novo Nordisk foundation

Participant: Pierre-Alexandre Mattei
External participants: Wouter Boomsma (University of Copenhagen), Jes Frellsen (Technical University of Denmark), Søren Hauberg (principal investigator, Technical University of Denmark), and Ole Winther (Technical University of Denmark)
Value: 180000 DKK ($\approx$ 24000 EUR)

The object of this grant is the organisation of the 2nd Copenhagen Workshop on Generative Models and Uncertainty Quantification (GenU) in 2021.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Inria international partners

Informal international partners The Maasai team has informal relationships with the following international teams:

- Department of Statistics of the University of Washington, Seattle (USA) through collaborations with Elena Erosheva and Adrian Raftery,
- SAILAB team at Universita di Sienna, Sienna (Italy) through collaborations with Marco Gori (details given below)
- School of Mathematics and Statistics, University College Dublin (Ireland) through the collaborations with Brendan Murphy, Riccardo Rastelli and Michael Fop,
- Department of Computer Science, University of Tübingen (Germany) through the collaboration with Ulrike von Luxburg,
- Université Laval, Québec (Canada) through the Research Program DEEL (DEpendable and Explainable Learning) with François Laviolette and Christian Gagné, and through a FFCR funding with Arnaud Droit (including the planned supervision of two PhD students in 2021), (details given below)
- DTU Compute, Technical University of Denmark, Copenhagen (Denmark), through collaborations with Jes Frellsen and his team (including the co-supervision of a PhD student in Denmark).

9.1.2 Participation in other international programs

DEpendable Explainable Learning Program (DEEL), Québec, Canada **Participants:** Frederic Precioso
Collaborations: François Laviolette (Prof. U. Laval), Christian Gagné (Prof. U. Laval)

The DEEL Project involves academic and industrial partners in the development of dependable, robust, explainable and certifiable artificial intelligence technological bricks applied to critical systems. We are involved in the Workpackage Robustness and the Workpackage Interpretability, in the co-supervision of several PhD thesis, Post-docs, and Master internships.

CHU Québec–Laval University Research Centre, Québec, Canada **Participants:** Pierre-Alexandre Mattei, Frederic Precioso, Louis Ohl (doctorant)
Collaborations: Arnaud Droit (Prof. U. Laval), Mickael Leclercq (Chercheur U. Laval), Khawla Seddiki (doctorante, U. Laval)

This collaboration framework covers several research projects: one project is related to the PhD thesis of Khawla Seddiki who works on Machine Learning/Deep Learning methods for classification and analysis of mass spectrometry data; another project is related to the France Canada Research Fund (FCRF) which provides the PhD funding of Louis Ohl, co-supervised by all the collaborators. This project investigates Machine Learning solutions for Aortic Stenosis (AS) diagnosis.

SAILAB: Lifelong learning in computer vision **Participants:** Lucile Sassatelli and Frédéric Precioso (UCA)
Keywords: computer vision, lifelong learning, focus of attention in vision, virtual video environments.
Collaborations: Dario (Universität Erlangen-Nürnberg), Alessandro Betti (UNISI), Stefano Melacci (UNISI), Matteo Tiezzi (UNISI), Enrico Meloni (UNISI), Simone Marullo (UNISI).

This collaboration concerns the current hot machine learning topics of Lifelong Learning, "on developing versatile systems that accumulate and refine their knowledge over time"), or continuous learning which targets tackling catastrophic forgetting via model adaptation. The most important expectations of this research is that of achieving object recognition visual skills by a little supervision, thus overcoming the need for the expensive accumulation of huge labelled image databases.

SAILAB: Streaming Virtual Reality: Learning for Attentional Models and Network Optimization **Participants:** Miguel Fabián Romero Rondón, Frédéric Precioso
Keywords: Modeling and prediction, Virtual Reality, Deep Learning, Multimedia Streaming
Collaborations: Lucile Sassatelli (Project Leader), Ramon Aparicio-Pardo, from I3S Lab of Université Côte d'Azur, and Dario Zanca and Marco Gori, University of Siena (to define gravitational laws to model attention in VR videos).

The goal of the project is to optimally decide what to transmit from a video sphere of VR content to maximize the streaming quality as observed in Fig. 24. We have made several contributions in 2020 [37, 18].

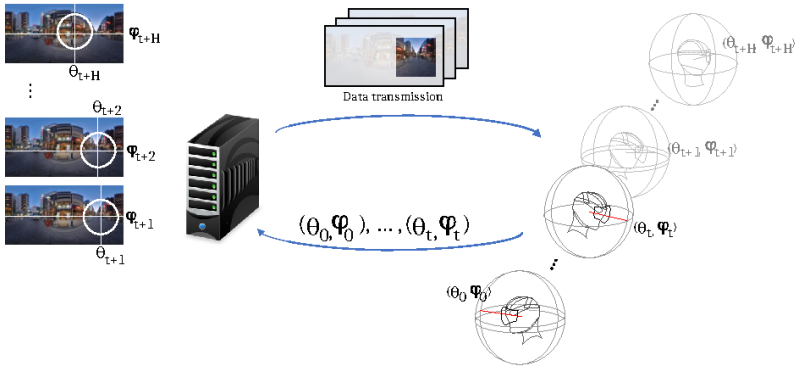


Figure 24: **Spherical video streaming principle**. The user requests the next video segment at time t, if the future orientations of the user $(\theta_{t+1},\varphi_{t+1}),...,(\theta_{t+H},\varphi_{t+H})$ were known, the bandwidth consumption could be reduced by sending in higher quality only the areas corresponding to the future FoV.

AI4Media: Audio-Video Similarity Learning **Participants:** Melissa Sanabria, Frédéric Precioso.
Collaborations: Ioannis Kompatsiaris. Centre for Research and Technology Hellas, Thermi-Thessaloniki, Greece.
Keywords: Multimodal Retrieval, Video Summarization, Multimodal data.

Creating new datasets is always a highly resource consuming task. Fortunately, with the current growth of video content available online, we have access to videos of entire events and their respective summaries. However, to create a ground truth dataset for summarization, we need the corresponding time intervals of each clip of the summary in the original video. For this reason, we have started a collaboration to explore audio-video similarity learning in order to create a summarization dataset.

9.2 International research visitors

9.2.1 Visits of international scientists

In 2020, we had the opportunity to welcome a few scientists, for several weeks each:

- Elena Erosheva: She is a full professor of Statistics at the University of Washington, Seattle (USA). She is also an International Chair in Data Science at Université Côte d'Azur, funded by the Idex Jedi program,
- Marco Gori: He is a full professor of Computer Science at Universita di Sienna, Sienna (Italy). He is also an International Chair of the Institut 3IA Côte d'Azur,
- Riccardo Rastelli: He is an assistant professor in Statistics at University College Dublin. His research interests include network analysis and Bayesian statistics.

9.3 European initiatives

9.3.1 FP7 & H2020 Projects

Maasai is one of the 3IA-UCA research teams of **AI4Media**, one of the 4 ICT-48 Center of Excellence in Artificial Intelligence which has started in September 2020. There are 30 partners (Universities and companies), and 3IA-UCA received about 325k€.

9.4 National initiatives

Institut 3IA Côte d'Azur Following the call of President Macron to found several national institutes in AI, we presented in front of a international jury our project for the Institut 3IA Côte d'Azur in April 2019. The project was selected for funding (50 M€ for the first 4 years, including 16 M€ from the PIA program) and started in september 2019. Charles Bouveyron and Marco Gori are two of the 29 3IA chairs which were selected *ab initio* by the international jury and is also acting in 2020 as the Deputy Scientific Director of the institute. The research of the institute is organized around 4 thematic axes: Core elements of AI, Computational Medicine, AI for Biology and Smart territories. The Maasai reserch team is totally aligned with the first axis of the Institut 3IA Côte d'Azur. The team has 4 Ph.D. students who are directly funded by the institute.
Web site: https://3ia.univ-cotedazur.eu

9.5 Regional initiatives

Centre de pharmacovigilance, CHU Nice **Participants:** Charles Bouveyron, Marco Corneli, Giulia Marchello, Michel Riveill, Xuchun Zhang
Keywords: Pharmacovigilance, co-clustering, count data, text data
Collaborateurs: Milou-Daniel Drici, Audrey Freysse, Fanny Serena Romani

The team works very closely with the Regional Pharmacovigilance Center of the University Hospital Center of Nice (CHU) through several projects. The first project concerns the construction of a dashboard to classify spontaneous patient and professional reports, but above all to report temporal breaks. To this end, we are studying the use of dynamic co-classification techniques to both detect significant ADR patterns and identify temporal breaks in the dynamics of the phenomenon. The second project focuses on the analysis of medical reports in order to extract, when present, the adverse events for characterization. After studying a supervised approach, we are studying techniques requiring fewer annotations.

Interpretability for automated decision services **Participants:** Damien Garreau, Frédéric Precioso
Keywords: interpretability, deep learning
Collaborations: Greger Ottosson (IBM)

Businesses rely more and more frequently on machine learning to make automated decisions. In addition to the complexity of these models, a decision is rarely by using only one model. Instead, the crude reality of business decision services is that of a jungle of models, each predicting key quantities for the problem at hand, that are then agglomerated to produce the final decision, for instance by a decision tree. In collaboration with IBM, we want to provide principled methods to obtain interpretability of these automated decision processes. An intern will start on the subject in March 2021.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

ICML Artemiss workshop Pierre-Alexandre Mattei co-founded and co-organized the first ICML workshop on the Art of learning with missing values (Artemiss). Artemiss was part of the 2020 edition of the International Conference on machine Learning (ICML).

The other founding co-organisers are Jes Frellsen (Technical University of Denmark), Julie Josse (Ecole Polytechnique and Inria), and Gaël Varoquaux (Inria).
Web site: https://artemiss-workshop.github.io/

52ème Journées de Statistique de la SFDS Charles Bouveyron, Marco Corneli, Damien Garreau, Pierre-Alexandre Mattei and Frédéric Precioso are members of the organization comittee of the conference "52ème Journées de Statistique", which is the annual conference of the French Statistical Association (SFdS). This conference usually gathers more than 400 statisticians from academic and industry. The 2020 edition was initially planned in Nice on 25-29 May 2020. Due to the pandemic, the conference has been

postponed by one year and will be held in Nice on 7-11 June 2021. Charles Bouveyron is the President of the organization comittee.
Web site: https://jds2021.sciencesconf.org

Statlearn workshop The workshop Statlearn is a scientific workshop held every year, which focuses on current and upcoming trends in Statistical Learning. Statlearn is a scientific event of the French Society of Statistics (SFdS) that has been organised since 2010. Conferences and tutorials are organized alternatively every other year. In 2020, a one-week spring-school should have been be held in Cargèse (March, 23-27), but has been postponed in 2022 due to the pandemic. The 2022 edition will be the 11th edition of the Statlearn series. The Statlearn conference was founded by Charles Bouveyron, and Marco Corneli, Damien Garreau and Pierre-Alexandre Mattei are members of the scientific comittee.
Web site: https://statlearn.sciencesconf.org

10.1.2 Scientific events: selection

Member of the conference program committees

- Frédéric Precioso is a member of the program committee of the the conference "52ème Journées de Statistique", which is the annual conference of the French Statistical Association (SFdS). The 2020 edition was initially planned in Nice on 25-29 May 2020. Due to the pandemic, the conference has been postponed by one year and will be held in Nice on 7-11 June 2021.

Member of the editorial boards

- Charles Bouveyron is Associate Editor for the Annals of Applied Statistics since 2016.

Reviewer - reviewing activities All permanent members of the team are serving as reviewers for the most important journals and conferences in statistical and machine learning, such as (non exhaustive list):

- International journals:
  - Annals of Applied Statistics,
  - Statistics and Computing,
  - Journal of the Royal Statistical Society, Series C,
  - Journal of Computational and Graphical Statistics,
  - Journal of Machine Learning Research
- International conferences:
  - Neural Information Processing Systems (Neurips),
  - International Conference on Machine Learning (ICML),
  - International Conference on Learning Representations (ICLR),
  - International Joint Conference on Artificial Intelligence (IJCAI),
  - International Conference on Artificial Intelligence and Statistics (AISTATS),
  - International Conference on Computer Vision and Pattern Recognition

10.1.3 Invited talks

- Charles Bouveyron was invited for a keynote lecture at the European Conference on Data Analysis, Napoli, Italy, September 2020 (the conference was cancelled due to the Covid pandemy).
- Pierre-Alexandre Mattei gave a talk at the online One World ABC seminar (video on YouTube).

10.1.4 Leadership within the scientific community

- Charles Bouveyron has been the Deputy Scientific Director of the Institut 3IA Côte d'Azur from September 2019 until December 2020.

10.1.5 Scientific expertise

Most permanent members of the team are serving as experts for the ANR or foreign research agencies.

10.1.6 Research administration

- Frédéric Precioso is the Scientific Responsible for AI at the French Research Agency (ANR) since September 2011.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

C. Bouveyron, D. Garreau, F. Precioso and M. Riveill are professors at Université Côte d'Azur and therefore handle usual teaching duties. M. Corneli and P.A. Mattei are also teaching around 60h per year at Université Côte d'Azur.

C. Bouveyron (up to august 2020) and M. Riveill (since September 2020) are responsible for the MSc. Data Sciences and Artificial Intelligence at Université Côte d'Azur.

10.2.2 Supervision

PhD students, postdocs, and interns of the team are listed in Section 1. Additionally, members of the team supervise several Masters projects, in particular from the MSc. Data Sciences and Artificial Intelligence at Université Côte d'Azur.

10.3 Popularization

10.3.1 Interventions

- Frederic Precioso has developed an experimental platform both for research projects and scientific mediation on the topic of autonomous cars. This platform is currently installed in the "Maison de l'Intelligence Artificielle" where high school students have already experimented coding autonomous remote control cars (https://maison-intelligence-artificielle.com/experimenter-projets-ia/).
- Charles Bouveyron has developed an interactive software allowing to visualise the relationships between pollution and a health disease (dispnea) in the Région Sud. This platform is currently installed in the "Maison de l'Intelligence Artificielle".

11 Scientific production

11.1 Major publications

[1] M. Corneli, C. Bouveyron and P. Latouche. 'Co-clustering of ordinal data via latent continuous random variables and not missing at random entries'. In: *Journal of Computational and Graphical Statistics* (2020). DOI: 10.1080/10618600.2020.1739533. URL: https://hal.archives-ouvertes.fr/hal-01978174.

[2] D. Garreau and U. Von Luxburg. 'Explaining the Explainer: A First Theoretical Analysis of LIME'. In: *AISTATS 2020 - 23rd International Conference on Artificial Intelligence and Statistics*. Palermo /Online, Italy, Aug. 2020. URL: https://hal.archives-ouvertes.fr/hal-02935171.

[3] N. B. Ipsen, P.-A. Mattei and J. Frellsen. 'not-MIWAE: Deep Generative Modelling with Missing not at Random Data'. In: *International Conference on Learning Representations*. Virtual conference (formerly planned in Vienna), Austria, 2021. URL: https://hal.inria.fr/hal-03044124.

[4] K. Seddiki, P. Saudemont, F. Precioso, N. Ogrinc, M. Wisztorski, M. Salzet, I. Fournier and A. Droit. 'Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification'. In: *Nature Communications* 11.1 (Dec. 2020). DOI: 10.1038/s41467-020-19354-z. URL: https://hal.archives-ouvertes.fr/hal-03132326.

11.2 Publications of the year

International journals

[5] S. Allassonnière and J. Chevallier. 'A New Class of EM Algorithms. Escaping Local Maxima and Handling Intractable Sampling'. In: *Computational Statistics and Data Analysis* (11th Dec. 2020). URL: https://hal.archives-ouvertes.fr/hal-02044722.

[6] C. Bouveyron, P. Latouche and P.-A. Mattei. 'Exact Dimensionality Selection for Bayesian PCA'. In: *Scandinavian Journal of Statistics* (2020). DOI: 10.1111/sjos.12424. URL: https://hal.archives-ouvertes.fr/hal-01484099.

[7] J. Chevallier, V. Debavelaere and S. Allassonnière. 'A coherent framework for learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data'. In: *SIAM Journal on Imaging Sciences* (2021). URL: https://hal.archives-ouvertes.fr/hal-01646298.

[8] M. Corneli, C. Bouveyron and P. Latouche. 'Co-clustering of ordinal data via latent continuous random variables and not missing at random entries'. In: *Journal of Computational and Graphical Statistics* (2020). DOI: 10.1080/10618600.2020.1739533. URL: https://hal.archives-ouvertes.fr/hal-01978174.

[9] E. Erosheva, S. Grant, M.-C. Chen, M. Lindner, R. Nakamura and C. Lee. 'NIH peer review: Criterion scores completely account for racial disparities in overall impact scores'. In: *Science Advances* 6.23 (3rd June 2020), eaaz4868. DOI: 10.1126/sciadv.aaz4868. URL: https://hal.archives-ouvertes.fr/hal-02870751.

[10] N. Jouvin, P. Latouche, C. Bouveyron, G. Bataillon and A. Livartowski. 'Greedy clustering of count data through a mixture of multinomial PCA'. In: *Computational Statistics* (2020). DOI: 10.1007/s00180-020-01008-9. URL: https://hal.archives-ouvertes.fr/hal-02278224.

[11] N. Keriven, D. Garreau and I. Poli. 'NEWMA: a new method for scalable model-free online change-point detection'. In: *IEEE Transactions on Signal Processing* 68 (27th Apr. 2020), pp. 3515–3528. DOI: 10.1109/TSP.2020.2990597. URL: https://hal.archives-ouvertes.fr/hal-02484988.

[12] A. Saint-Dizier, J. Delon and C. Bouveyron. 'A unified view on patch aggregation'. In: *Journal of Mathematical Imaging and Vision* (2020). DOI: 10.1007/s10851-019-00921-z. URL: https://hal.archives-ouvertes.fr/hal-01865340.

[13] A. Schmutz, J. Jacques, C. Bouveyron, L. Cheze and P. Martin. 'Clustering multivariate functional data in group-specific functional subspaces'. In: *Computational Statistics* (12th Feb. 2020). DOI: 10.1007/s00180-020-00958-4. URL: https://hal.inria.fr/hal-01652467.

[14] K. Seddiki, P. Saudemont, F. Precioso, N. Ogrinc, M. Wisztorski, M. Salzet, I. Fournier and A. Droit. 'Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification'. In: *Nature Communications* 11.1 (Dec. 2020). DOI: 10.1038/s41467-020-19354-z. URL: https://hal.archives-ouvertes.fr/hal-03132326.

[15] T. U. Tran, H. T. Ti Hoang, P. Hoai Dang and M. Riveill. 'Multitask Aspect_Based Sentiment Analysis with Integrated Bidirectional LSTM & CNN Model'. In: *IAES International Journal of Artificial Intelligence (IJ-AI)* 9.1 (1st Mar. 2020). DOI: 10.11591/ijai.v9.i1. URL: https://hal.archives-ouvertes.fr/hal-03052803.

[16] D. Zanca, M. Gori, S. Melacci and A. Rufa. 'Gravitational Models Explain Shifts on Human Visual Attention'. In: *Scientific Reports* (2020). URL: https://hal.archives-ouvertes.fr/hal-02940854.

International peer-reviewed conferences

[17] N. B. Ipsen, P.-A. Mattei and J. Frellsen. 'not-MIWAE: Deep Generative Modelling with Missing not at Random Data'. In: International Conference on Learning Representations. Virtual conference (formerly planned in Vienna), Austria: https://openreview.net/forum?id=tu29GQT0JFy, 2021. URL: https://hal.inria.fr/hal-03044124.

[18] M. F. Romero Rondon, L. Sassatelli, R. Aparicio-Pardo and F. Precioso. 'A Unified Evaluation Framework for Head Motion Prediction Methods in 360° Videos'. In: MMSys '20 - 11th ACM Multimedia Systems Conference. Vol. 20. MMSys '20: Proceedings of the 11th ACM Multimedia Systems Conference. Istanbul, Turkey, 8th June 2020, pp. 279–284. DOI: 10.1145/3339825.3394934. URL: https://hal.archives-ouvertes.fr/hal-02615979.

[19] M. Sanabria, F. Precioso and T. Menguy. 'Profiling Actions for Sport Video Summarization: An attention signal analysis'. In: MMSP 2020 - 22nd IEEE International Workshop on Multimedia Signal Processing. Tampere, Finland, 21st Sept. 2020. URL: https://hal.inria.fr/hal-02910211.

[20] T. Yang, N. Pasquier, A. Hom, L. Dollé and F. Precioso. 'Semi-supervised Consensus Clustering Based on Frequent Closed Itemsets: Amadeus Intellectual Property Invention Patent ID2326WW00 "Clustering Techniques for Revenue Accounting Error-Handling Automation" Defensive Paper'. In: CIKM'2020 29th ACM International Conference on Information and Knowledge Management (Acceptance Rate: 18%). Proceedings of the CIKM'2020 29th ACM International Conference on Information and Knowledge Management. Galway, Ireland: https://www.cikm2020.org/, 19th Oct. 2020, pp. 3341–3344. DOI: 10.1145/3340531.3417453. URL: https://hal.archives-ouvertes.fr/hal-02917863.

[21] T. Yang, N. Pasquier and F. Precioso. 'Ensemble Clustering Based Semi-Supervised Learning for Revenue Accounting Workflow Management: Amadeus Intellectual Property Invention Patent ID2326WW00 "Clustering Techniques for Revenue Accounting Error-Handling Automation" Defensive Paper'. In: DATA'2020 International Conference on Data Science, Technology and Applications (Acceptance Rate: 14%). Proceedings of the DATA'2020 International Conference on Data Science, Technology and Applications. Paris, France: http://www.dataconference.org, 7th July 2020, pp. 283–293. DOI: 10.5220/0009883802830293. URL: https://hal.archives-ouvertes.fr/hal-02540832.

Conferences without proceedings

[22] G. Ciravegna, F. Giannini, M. Gori, M. Maggini and S. Melacci. 'Human-Driven FOL Explanations of Deep Learning'. In: IJCAI-PRICAI 2020 - 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence. Yokohama, Japan, 11th July 2020, pp. 2234–2240. DOI: 10.24963/ijcai.2020/309. URL: https://hal.archives-ouvertes.fr/hal-03045280.

[23] J. A. García, F. Precioso, P. Staccini and M. Riveill. 'DiagnoseNET: Automatic Framework to Scale Neural Networks on Heterogeneous Systems Applied to Medical Diagnosis'. In: ICITCS 2020 - 8th International Conference on IT Convergence and Security. Nha Trang / Virtual, Vietnam, 19th Aug. 2020. URL: https://hal.archives-ouvertes.fr/hal-02869960.

[24] D. Garreau and U. Von Luxburg. 'Explaining the Explainer: A First Theoretical Analysis of LIME'. In: AISTATS 2020 - 23rd International Conference on Artificial Intelligence and Statistics. Palermo /Online, Italy, 26th Aug. 2020. URL: https://hal.archives-ouvertes.fr/hal-02935171.

[25] W. Hao, C. Dartigues-Pallez and M. Riveill. 'Supervised learning for Human Action Recognition from multiple Kinects'. In: BDMS 2020 - 7th Big Data Management and Service in DASFAA 2020. Jeju, South Korea, 24th Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-02869941.

[26] N. B. Ipsen, P.-A. Mattei and J. Frellsen. 'How to deal with missing data in supervised deep learning?' In: Artemiss - ICML Workshop on the Art of Learning with Missing Values. Vienne, Austria: https://artemiss-workshop.github.io/, July 2020. URL: https://hal.inria.fr/hal-03044144.

[27] L. C. Lamb, A. Garcez, M. Gori, M. Prates, P. Avelar and M. Vardi. 'Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective'. In: IJCAI-PRICAI 2020 - 29th International Joint Conference on Artificial Intelligence - Pacific Rim International Conference on Artificial Intelligence. Yokohama, Japan, 11th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02878531.

[28] D. Liang, M. Corneli, P. Latouche and C. Bouveyron. 'Missing rating imputation based on product reviews via deep latent variable models'. In: ICML2020 Workshop on the Art of Learning with Missing Values (Artemiss). Virtual, France, 17th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02933326.

[29] G. Marchello, M. Corneli and C. Bouveyron. 'The Dynamic Latent Block Model for Sparse and Evolving Count Matrices'. In: ICML Workskshop Artemiss. Nice, France, 17th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02971127.

[30] G. Marchello, M. Corneli and C. Bouveyron. 'The dynamic latent block model for the co-clustering of evolving binary matrices'. In: SFdS 2020 - 52èmes journées de Statistique de la la Société Française de Statistique. Nice, France, 7th June 2020. URL: https://hal.archives-ouvertes.fr/hal-02972985.

[31] P.-A. Mattei and J. Frellsen. 'Negative Dependence Tightens Variational Bounds'. In: ICML 2020 - 2nd Workshop on Negative Dependence and Submodularity for ML. Vienne / Online, Austria, 17th July 2020. URL: https://hal.inria.fr/hal-03044115.

[32] D. Mayaffre, C. Bouzereau, M. Guaresi, F. Precioso and L. Vanni. 'Du texte à l'intertexte. Le palimpseste Macron au révélateur de l'Intelligence artificielle'. In: CMLF 2020 - 7ème Congrès mondiale de linguistique française. Montpellier / Online, France, 6th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02520224.

[33] D. Mayaffre and L. Vanni. 'Objectiver l'intertexte ? Emmanuel Macron, deep learning et statistique textuelle'. In: JADT 2020 - 15èmes Journées Internationales d'Analyse statistique des Données Textuelles. Toulouse, France, 16th June 2020. URL: https://hal.archives-ouvertes.fr/hal-02894990.

[34] E. Meloni, L. Pasqualini, M. Tiezzi, M. Gori and S. Melacci. 'SAILenv: Learning in Virtual Visual Environments Made Simple'. In: 25th International Conference on Pattern Recognition. Milan, Italy, 10th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-02965715.

[35] F. Orlhac, T. Cassou-Mounat, J.-Y. Pierga, M. Luporsi, C. Nioche, C. Bouveyron, N. Ayache, N. Jehanno, A. Livartowski and I. Buvat. 'Can we identify "twin patients" to predict response to neoadjuvant chemotherapy in breast cancer?' In: SNMMI Annual Meeting. Virtual Meeting, United States, 11th July 2020. URL: https://www.hal.inserm.fr/inserm-02952453.

[36] F. Orlhac, A.-C. Rollet, I. Buvat, J. Darcourt, V. Bourg, C. Nioche, C. Bouveyron, N. Ayache and O. Humbert. 'Identifying a reliable radiomic signature from scarce data: illustration for 18F-FDOPA PET images in glioblastoma patients'. In: EANM Annual Meeting - Annual Meeting of the European Association of Nuclear Medicine. Virtual Meeting, Austria, 22nd Oct. 2020. URL: https://www.hal.inserm.fr/inserm-02952445.

[37] M. F. Romero Rondon, L. Sassatelli, R. Aparicio-Pardo and F. Precioso. 'TRACK: A Multi-Modal Deep Architecture for Head Motion Prediction in 360-Degree Videos'. In: IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates, 25th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02615980.

[38] M. Sanabria, F. Precioso and T. Menguy. 'Hierarchical Multimodal Attention for Deep Video Summarization'. In: 25th International Conference on Pattern Recognition. Milan, Italy, 10th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-02964209.

[39] M. Tiezzi, S. Melacci, A. Betti, M. Maggini and M. Gori. 'Focus of Attention Improves Information Transfer in Visual Features'. In: NeurIPS 2020 - 34th Conference on Neural Information Processing Systems. Vancouver / Online, Canada, 6th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-02878372.

[40] L. Vanni, M. Corneli, D. Longrée, D. Mayaffre and F. Precioso. 'Hyperdeep : deep learning descriptif pour l'analyse de données textuelles'. In: JADT 2020 - 15èmes Journées Internationales d'Analyse statistique des Données Textuelles. Toulouse, France, 16th June 2020. URL: https://hal.archives-ouvertes.fr/hal-02926880.

Scientific book chapters

[41] C. Bouveyron. 'High-Dimensional Statistical Learning and Its Application to Oncological Diagnosis by Radiomics'. In: *Healthcare and Artificial Intelligence*. 18th Mar. 2020, pp. 121–128. DOI: 10.1007/978-3-030-32161-1_17. URL: https://hal.archives-ouvertes.fr/hal-02516907.

[42] Y. Li, F. Precioso and C. Liu. 'Visualizing ECG Contribution into Convolutional Neural Network Classification'. In: *Feature Engineering and Computational Intelligence in ECG Monitoring*. 25th June 2020, pp. 157–174. DOI: 10.1007/978-981-15-3824-7_9. URL: https://hal.archives-ouvertes.fr/hal-03132410.

[43] L. Vanni, M. Corneli, D. Longrée, D. Mayaffre and F. Precioso. 'Key Passages : From statistics to Deep Learning'. In: *Text Analytics. Advances and Challenges*. 2020, pp. 41–54. DOI: 10.1007/978-3-030-52680-1_4. URL: https://hal.archives-ouvertes.fr/hal-03099658.

Reports & preprints

[44] C. Bouveyron, J. Jacques, A. Schmutz, F. Simoes and S. Bottini. *Co-Clustering of Multivariate Functional Data for the Analysis of Air Pollution in the South of France*. 1st June 2020. URL: https://hal.archives-ouvertes.fr/hal-02862177.

[45] E. Côme, P. Latouche, N. Jouvin and C. Bouveyron. *Hierarchical clustering with discrete latent variable models and the integrated classification likelihood*. 3rd Apr. 2020. URL: https://hal.archives-ouvertes.fr/hal-02530705.

[46] M. Corneli and E. Erosheva. *A Bayesian approach for clustering and exact finite-sample model selection in longitudinal data mixtures*. 9th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02310069.

[47] S. Daga, C. Fallerini, M. Baldassarri, F. Fava, F. Valentino, G. Doddato, E. Benetti, S. Furini, A. Giliberti, R. Tita, S. Amitrano, M. Bruttini, I. Meloni, A. M. Pinto, F. Raimondi, A. Stella, F. Biscarini, N. Picchiotti, M. Gori, P. Pinoli, S. Ceri, M. Sanarico, F. Crawley, A. Renieri, F. Mari and E. Frullanti. *Employing a Systematic Approach to Biobanking and Analyzing Genetic and Clinical Data for Advancing COVID-19 Research*. 7th Dec. 2020. DOI: 10.1101/2020.07.24.20161307. URL: https://hal.archives-ouvertes.fr/hal-03045235.

[48] L. Faggi, A. Betti, D. Zanca, S. Melacci and M. Gori. *Wave Propagation of Visual Stimuli in Focus of Attention*. 23rd June 2020. URL: https://hal.archives-ouvertes.fr/hal-02878300.

[49] M. Fop, P.-A. Mattei, C. Bouveyron and T. B. Murphy. *Unobserved classes and extra variables in high-dimensional discriminant analysis*. 5th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03132362.

[50] D. Garreau and U. Von Luxburg. *Looking Deeper into Tabular LIME*. 24th Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-02948641.

[51] M. Gori. *An Overview on the Web of Clinical Data*. 16th Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-02940789.

[52] M. Gori and A. Betti. *Backprop Diffusion is Biologically Plausible*. 23rd June 2020. URL: https://hal.archives-ouvertes.fr/hal-02878574.

[53] N. Jouvin, C. Bouveyron and P. Latouche. *A Bayesian Fisher-EM algorithm for discriminative Gaussian subspace clustering*. 9th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03047930.

[54] D. Liang, M. Corneli, C. Bouveyron and P. Latouche. *DeepLTRS: A Deep Latent Recommender System based on User Ratings and Reviews*. 24th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03021362.

[55] G. Marchello, A. Fresse, M. Corneli and C. Bouveyron. *Co-clustering of evolving count matrices in pharmacovigilance with the dynamic latent block model*. 19th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03146769.

[56] D. Mardaoui and D. Garreau. *An Analysis of LIME for Text Data*. 26th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02977786.

[57] P.-A. Mattei. *A Parsimonious Tour of Bayesian Model Uncertainty*. 7th Dec. 2020. URL: https://hal.inria.fr/hal-03044295.

[58] C. Vincent-Cuaz, T. Vayer, R. Flamary, M. Corneli and N. Courty. *Online Graph Dictionary Learning*. 12th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03140349.

[59] A. Zugarini, E. Meloni, A. Betti, A. Panizza, M. Corneli and M. Gori. *An Optimal Control Approach to Learning in SIDARTHE Epidemic model*. 7th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03044564.

Other scientific publications

[60] T. Yang, N. Pasquier and F. Precioso. *Semi-Supervised Consensus Clustering Based on Frequent Closed Itemsets*. Bodenseeforum, Germany, 27th Apr. 2020. URL: https://hal.archives-ouvertes.fr/hal-02540836.

11.3 Other

Patents

[61] T. Yang, N. Pasquier, F. Precioso, A. Hom and L. Dollé. 'Clustering Techniques for Revenue Accounting Error-Handling Automation'. 7th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02926737.